

A genome-wide association study in chronic thromboembolic pulmonary hypertension and the ADAMTS13-VWF axis



Michael Newnham

Supervisors:

Prof. Nicholas Morrell

Dr. Mark Toshner

Department of Medicine

University of Cambridge

This degree is submitted for the degree of
Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the Degree Committee for the Faculties of Clinical Medicine and Veterinary Medicine.

Summary

A genome-wide association study in chronic thromboembolic pulmonary hypertension and the ADAMTS13-VWF axis

Michael Newnham

Chronic thromboembolic pulmonary hypertension (CTEPH) is an important and severe consequence of pulmonary embolism (PE), resulting from failure of thrombus resolution. Identifying genetic risk factors for CTEPH would provide important insights into pathobiology and might allow risk-stratification following PE. A genome-wide association study (GWAS) was performed in 1250 CTEPH patients, 1492 healthy controls and ~7 million single-nucleotide polymorphisms to identify novel disease loci.

The *ABO* locus was identified as the most significant common variant genetic association with CTEPH in both a discovery and validation cohort. The A1 subgroup of *ABO* was enriched in CTEPH and this may result in multiple functional consequences including variation in plasma von Willebrand factor (VWF) levels.

Abnormalities in haemostasis are implicated in CTEPH pathobiology, including elevated levels of VWF, which is cleaved by ADAMTS13 (a disintegrin and metalloproteinase with a thrombospondin type 1 motif, member 13). The ADAMTS13-VWF axis was investigated in 208 CTEPH patients including its relationship to ABO blood groups and *ADAMTS13* genetic variants.

Plasma ADAMTS13 levels are markedly reduced in CTEPH. This is independent of pulmonary hypertension, disease severity or systemic inflammation. Plasma VWF levels were confirmed to be markedly increased in CTEPH. These findings implicate dysregulation of the ADAMTS13-VWF axis in CTEPH pathobiology.

Preface

I would like to thank Dr Mark Toshner, Professor Nicholas Morrell and Dr Joanna Pepke-Zaba for setting up the study and recruiting participating specialist pulmonary hypertension centres. I would particularly like to thank them for giving me the opportunity to undertake such interesting and challenging research at both Royal Papworth Hospital and the University of Cambridge. They have provided encouragement, mentorship and guidance during my PhD. The funding and support they supplied were critical for enabling me to undertake this work. I am also grateful to the British Heart Foundation for providing funding for 1 year. I would like to acknowledge the help of all the pulmonary hypertension centres, research nurses and clinical staff involved in the recruitment of patients, particularly the pulmonary vascular research team at Royal Papworth Hospital. I thank the patients who were recruited to this study.

I am grateful to a number of people that I collaborated with during my PhD. Dr Marta Bleda (University of Cambridge) performed the initial pilot GWAS analysis in 2014, power calculations and genotype clustering. Dr Stefan Gräf's group provided additional help and support with the computational biological elements of the study. Professor David Lane and Professor Mike Laffan (Imperial College London) supplied the ADAMTS13 / VWF antibodies and Dr Kieron South (Imperial College London) assisted in performing the laboratory work. Professor Martin Wilkins and Dr John Wharton (Imperial College London) supplied some healthy control and pulmonary embolism blood samples for the study. Dr Mark Southwood (Royal Papworth Hospital) assisted with the immunohistochemistry.

My parents have always supported me, and their lifelong hard work has given me many opportunities that they did not have. I am forever grateful for their love and support.

Some of this work has been published in a peer reviewed international journal:

Newnham M, South K, Bleda M, Auger WR, Barbera JA, Bogaard H, et al. The ADAMTS13-VWF axis is dysregulated in chronic thromboembolic pulmonary hypertension. Eur Respir J. 2019;53(3) © ERS 2019

Contents

Declaration	i
Summary	ii
Preface	iii
Contents	iv
Figures	x
Tables	xiii
Abbreviations	xv

1 Introduction	1
1.1 Literature search	1
1.2 Chronic thromboembolic pulmonary hypertension	1
1.2.1 Overview	1
1.2.1.1 CTEPH management	4
1.2.2 CTEPH pathobiology	5
1.2.2.1 CTEPH histology	5
1.2.2.2 CTEPH pathobiology: introduction	6
1.2.2.3 Increased thrombus	6
1.2.2.4 Fibrinolysis	7
1.2.2.5 Neoangiogenesis	9
1.2.2.6 Inflammation and endothelial dysfunction	9
1.2.2.7 Problems investigating CTEPH pathobiology	9
1.3 Genetics of CTEPH	10
1.3.1 Traditional thrombophilia risk factors	10
1.3.1.1 Heritable thrombophilia overview	10
1.3.1.2 Heritable thrombophilias and CTEPH	13
1.3.2 ABO and fibrinogen genes	14
1.3.3 Genetic variants associated with pulmonary arterial hypertension in CTEPH	15
1.3.3.1 Pulmonary arterial hypertension overview	15
1.3.3.2 Pulmonary arterial hypertension genetic variants and CTEPH	18
1.3.3.3 Other CTEPH genetic associations	19

1.3.4	CTEPH genetic architecture	19
1.3.5	CTEPH epigenetics	21
1.3.6	Genetics of CTEPH: summary.....	21
1.4	Genetics of venous thromboembolism	22
1.4.1	Traditional thrombophilia risk factors	22
1.4.2	Venous thromboembolism and <i>ABO</i>	24
1.5	Genome-wide association studies.....	24
1.5.1	GWAS background	25
1.5.2	GWAS methods	29
1.5.3	GWAS association testing and statistics	33
1.5.4	GWAS challenges.....	39
1.5.5	Venous thromboembolism GWASs	40
1.6	Von Willebrand Factor and ADAMTS13.....	42
1.6.1	Thrombotic thrombocytopenic purpura	43
1.6.2	VWF, ADAMTS13 and thrombotic diseases.....	44
1.7	Pilot CTEPH GWAS data	45
1.8	Hypotheses and Aims	46
2	Materials and Methods	47
2.1	GWAS	47
2.1.1	Sample size calculations	47
2.1.2	Study samples and participants.....	47
2.1.3	DNA extraction and DNA microarray	49
2.1.4	GWAS quality control.....	49
2.1.4.1	GWAS quality control: overview	49
2.1.4.2	Micro-array intensity data quality control and genotype calling	50
2.1.4.3	Sample quality control	51
2.1.4.4	SNP quality control	52
2.1.5	Phasing, genetic imputation and post-imputation SNP QC	52
2.1.6	Association testing.....	53
2.1.7	Linkage disequilibrium	53
2.1.8	Genetic <i>ABO</i> groups.....	54
2.1.9	Fine mapping	55

2.1.9.1	99% credible set	55
2.1.9.2	Genomic functional annotations	55
2.2	ADAMTS13-VWF axis	56
2.2.1	Study samples and participants	56
2.2.2	ADAMTS13 and VWF plasma concentrations	57
2.2.2.1	ADAMTS13 plasma concentration	58
2.2.2.2	VWF plasma concentration	59
2.2.2.3	Replicate sample measurements	59
2.2.3	ADAMTS13 activity, D-dimer and VWF multimeric size	59
2.2.3.1	ADAMTS13 activity	60
2.2.3.2	D-Dimer plasma levels	60
2.2.3.3	VWF multimeric size	60
2.2.4	Immunohistochemistry	61
2.2.5	Protein quantitative trait loci	62
2.2.6	Clinical phenotype data	63
2.2.7	Statistical analysis	63
2.3	CTEPH phenotype–genotype associations	65
2.3.1	Phenotype data	65
2.3.1.1	Data extraction and quality control steps	65
2.3.1.2	Data centralisation	68
2.3.2	GWAS associations	70
2.3.2.1	Additional case-control analysis	70
2.3.2.2	Additional phenotype-genotype associations	71
2.4	Software and online tools	72
3	GWAS	73
3.1	Introduction	73
3.2	Results	74
3.2.1	Study samples and participants	74
3.2.2	Study exclusions and GWAS quality control	75
3.2.2.1	Sample exclusions: overview	75
3.2.2.2	Sample exclusions: GWAS quality control	77
3.2.2.3	SNP exclusions: GWAS quality control	83
3.2.2.3.1	SNP exclusions: overview	83

3.2.2.3.2	Micro-array clustering, genotype calling and exclusions	84
3.2.2.3.3	SNP genotype missingness, deviations from Hardy-Weinberg distribution and pre-imputation minor allele frequency	87
3.2.2.3.4	Post GWAS QC: minor allele frequency and imputation quality	90
3.2.2.4	Residual population structure	92
3.2.2.5	Study participant characteristics post-QC	95
3.2.3	GWAS statistical association testing	96
3.2.3.1	Joint analysis: discovery and validation cohorts combined	96
3.2.3.1.1	Association testing without covariates	96
3.2.3.1.2	Association testing adjusted for population stratification	98
3.2.3.1.3	Independent associations	106
3.2.3.2	Discovery cohort analysis	109
3.2.3.3	Validation cohort analysis	115
3.2.3.4	Genotyping quality of the significant GWAS associations	119
3.2.4	The <i>ABO</i> association	121
3.2.5	Fine mapping	126
3.2.5.1	Credible set analysis	126
3.2.5.2	Genomic functional annotation	128
3.2.6	Gene-based and gene-set analysis	132
3.2.7	GWAS putative associations	134
3.3	Discussion	136
3.3.1	Overview	136
3.3.2	Associated Loci	136
3.3.2.1	The <i>ABO</i> association	136
3.3.2.2	The <i>F11</i> putative association	137
3.3.3	Genomic functional annotations, gene-set and gene-based analyses	138
3.3.4	Absence of genetic associations	138
3.3.4.1	The <i>FGA-FGB-FGG</i> locus	139
3.3.4.2	Pulmonary hypertension related genes in CTEPH	140

3.3.5	Strengths and limitations	140
4	The ADAMTS13-VWF axis.....	143
4.1	Introduction.....	143
4.2	Results	144
4.2.1	Study samples and participants.....	144
4.2.2	ADAMTS13 and VWF plasma concentrations.....	146
4.2.2.1	ADAMTS13 plasma concentrations.....	146
4.2.2.2	VWF plasma concentrations.....	150
4.2.2.3	Interaction effects	155
4.2.2.4	Replicates and ADAMTS13 batch adjustment	157
4.2.2.5	ADAMTS13 and VWF: pre- and post-pulmonary endarterectomy.....	159
4.2.3	ADAMTS13 activity, D-dimer and VWF multimers	160
4.2.3.1	ADAMTS13 activity.....	160
4.2.3.2	D-dimers	160
4.2.3.3	VWF multimeric size.....	161
4.2.4	Clinical phenotype associations with ADAMTS13 and VWF	165
4.2.5	ADAMTS13-VWF and genotype analyses.....	171
4.2.5.1	Genetic <i>ABO</i> groups and ADAMTS13-VWF	171
4.2.5.2	Protein quantitative trait loci for ADAMTS13	174
4.2.6	Immunohistochemistry.....	176
4.3	Discussion	179
4.3.1	Overview.....	179
4.3.2	ADAMTS13-VWF plasma levels and other diseases	179
4.3.3	ADAMTS13-VWF: dysregulation mechanism and role in CTEPH pathobiology	180
4.3.4	ADAMTS13-VWF and <i>ABO</i>	181
4.3.5	ADAMTS13 protein quantitative trait loci in CTEPH	182
4.3.6	Strengths and limitations	182
5	CTEPH phenotype - genotype associations	184
5.1	Introduction.....	184
5.2	Results	186

5.2.1	Data capture, QC and missingness	186
5.2.2	Additional GWAS case-control analysis	189
5.2.2.1	<i>ABO</i> and <i>F11</i> risk alleles and CTEPH.....	189
5.2.2.2	Venous thromboembolism genes in CTEPH.....	190
5.2.2.3	SNPs associated with warfarin metabolism in the CTEPH GWAS.....	194
5.2.3	Additional phenotype-genotype associations	199
5.2.3.1	<i>ABO</i> and CTEPH disease severity	199
5.2.3.2	<i>ABO</i> groups and CTEPH survival.....	203
5.2.3.3	CTEPH disease distribution GWAS.....	204
5.2.3.4	CTEPH haemodynamics GWAS	207
5.3	Discussion	210
5.3.1	The effect of combining the <i>ABO</i> and <i>F11</i> risk alleles on CTEPH risk.....	210
5.3.2	The differential genetic associations between CTEPH, VTE and warfarin metabolism.....	210
5.3.3	The effect of genetic <i>ABO</i> groups on CTEPH disease severity and survival.....	211
5.3.4	CTEPH disease distribution and haemodynamics GWASs.....	212
6	Conclusions	214
7	Future studies	218
7.1	GWAS	218
7.2	ADAMTS13-VWF	221
7.3	Clinical perspectives.....	222
	References	224
	Appendix	243

Figures

1.1	Imaging modalities for diagnosing CTEPH.....	3
1.2	Pulmonary endarterectomy specimen	5
1.3	Blood coagulation overview	7
1.4	The fibrinolysis pathway	8
1.5	Activated protein C and the factor V Leiden mutation	11
1.6	Activation and Inhibition in the coagulation system	12
1.7	HPAH pathways and gene variants.....	17
1.8	The disease effect size of alleles with varying frequencies	20
1.9	GWAS SNP-trait associations have increased over time	27
1.10	SNP microarray overview	31
1.11	GWAS overview	32
1.12	Linkage and linkage disequilibrium.....	34
1.13	Genetic imputation.....	37
1.14	Manhattan plot of VTE associated genetic loci.....	41
1.15	Pathophysiology of thrombotic thrombocytopenic purpura.....	44
2.1	GWAS sample size calculations.....	48
2.2	Flow chart of GWAS analysis steps	50
2.3	Flow chart of study design and study participant numbers.....	57
2.4	Indirect immunohistochemistry using a polymer-based detection system.....	62
2.5	Flow chart of phenotype data extraction and quality control	66
2.6	Electronic case report form for right heart catheterisation data.....	69
3.1	GWAS sample numbers prior to exclusions	74
3.2	GWAS sample exclusions	76
3.3	Divergent ancestry assessed by principal component analysis	78
3.4	Sample relatedness assessed by identity by descent	80
3.5	Outlying sample heterozygosity plotted against sample genotype missingness.....	81
3.6	Discordant sex for individual samples	82
3.7	Total sample numbers following QC exclusions.....	83
3.8	Manhattan plot of all associations including incorrect genotype clustering and calling	85
3.9	Micro-array intensity clustering and false positive associations	86

3.10	Genotype missingness SNP exclusions	87
3.11	SNP Hardy-Weinberg equilibrium exclusions	88
3.12	SNP minor allele frequency distribution prior to imputation	89
3.13	SNP imputation quality	91
3.14	Minor allele frequency post imputation	91
3.15	Principal component analysis to detect residual population structure for each batch	93
3.16	Principal component analysis to detect residual population structure for combined batches	94
3.17	Case-control association testing without additional covariates: joint analysis..	97
3.18	Case-control association testing without covariates: discovery and validation cohorts	98
3.19	Case-control association testing with 5 ancestry informative principal components: joint analysis	100
3.20	Batch and centre association testing with CTEPH	106
3.21	Regional association plot of the associated locus in chromosome 9	107
3.22	Linkage disequilibrium heat maps of significant SNPs in the <i>ABO</i> and <i>ADAMTS13</i> loci	108
3.23	Conditional analysis at the associated chromosome 9 locus	109
3.24	Associated loci in the discovery cohort.....	110
3.25	Associated loci in the validation cohort.....	115
3.26	Case-control association testing pre-imputation.....	120
3.27	Micro-array clustering plots for the lead SNP associations	121
3.28	Genetic <i>ABO</i> groups in CTEPH and healthy controls	124
3.29	Genetic <i>ABO</i> groups and CTEPH risk.....	125
3.30	The percentage of genetic <i>ABO</i> groups in recruiting study centres	126
3.31	Regional association plot of 99% credible SNP set for the chromosome 9 association.....	128
3.32	Fine mapping: functional annotations for the 99% credible SNP set in chromosome 9.....	129
3.33	Circos plots of chromatin interactions and eQTLs for the associations in chromosome 9 (combined analysis) and chromosome 4 (discovery cohort) .	131
3.34	Gene-based association testing: combined group (discovery and validation cohort)	132

3.35	MAGMA tissue expression analysis	134
4.1	ADAMTS13 and VWF antigen (Ag) levels by diagnostic groups.....	148
4.2	Correlation of ADAMTS13 with VWF antigen levels in CTEPH and healthy controls	152
4.3	The odds ratios of CTEPH in relation to healthy controls for combined ADAMTS13 and VWF groups.....	153
4.4	ADAMTS13 multivariable linear model interaction effects.....	156
4.5	ADAMTS13 and VWF antigen levels by diagnostic groups for batch1.....	158
4.6	ADAMTS13 and VWF antigen levels pre- and post-pulmonary endarterectomy.....	159
4.7	ADAMTS13 activity, D-dimer and VWF multimeric size in CTEPH and healthy controls	162
4.8	ADAMTS13 activity, D-dimer and VWF multimeric size correlation in CTEPH and healthy controls	164
4.9	Correlation of ADAMTS13 and VWF antigen levels with markers of disease severity in CTEPH at baseline	166
4.10	Correlation of ADAMTS13 and VWF antigen levels with blood markers of inflammation at baseline	168
4.11	ADAMTS13 and VWF antigen levels in CTEPH sub-diagnostic and post-PEA residual pulmonary hypertension groups	169
4.12	ADAMTS13 and VWF antigen levels in PE stratified by residual perfusion defects and provoked PE	170
4.13	ADAMTS13 and VWF antigen levels by <i>ABO</i> genetic groups.....	172
4.14	ADAMTS13 and VWF antigen levels by comprehensive <i>ABO</i> genetic groups.....	173
4.15	Expression of ADAMTS13 in lung tissue evaluated by immunohistochemistry	177
5.1	Density plots of selected haemodynamics pre- and post-QC.....	188
5.2	<i>ABO</i> and <i>F11</i> risk alleles and CTEPH.....	189
5.3	VTE associated loci in the CTEPH GWAS	192
5.4	Power calculations for the <i>F5</i> VTE associated loci.....	195
5.5	Warfarin metabolism associated loci in the CTEPH GWAS	197
5.6	The effect of risk (effect) alleles for Chr4 (<i>F11</i>) and Chr9 (<i>ABO</i>) on haemodynamics	201

5.7	Genetic <i>ABO</i> groups and CTEPH disease severity.....	202
5.8	CTEPH survival following pulmonary endarterectomy in different <i>ABO</i> groups.....	203
5.9	CTEPH disease subtypes by recruiting centre	205
5.10	Distal/Proximal CTEPH GWAS association testing	206
5.11	CTEPH haemodynamic GWAS association testing	208

Tables

1.1	Heritable thrombophilias.....	13
1.2	Clinical classification of PAH (group 1 PH).....	16
1.3	Risk of VTE with different hereditary thrombophilias.....	23
1.4	GWAS genetic applications	29
2.1	Haplotypes used to reconstruct genetic <i>ABO</i> groups.....	54
2.2	Summary of extracted datasets.....	67
2.3	Phenotype domains and example eCRFs for OpenClinica data capture	70
3.1	Total number of SNP exclusions per batch prior to imputation	84
3.2	SNP exclusions from micro-array clustering quality thresholds that were applied without a re-clustering step	84
3.3	SNP exclusions for each quality control step	90
3.4	Baseline characteristics for the case-control groups included in association testing	95
3.5	Ancestry informative eigenvectors and case-control status	99
3.6	Significant SNPs in the joint analysis	101
3.7	Significant SNPs in the discovery cohort.....	111
3.8	Significant SNPs in the validation cohort.....	117
3.9	Significant SNPs in the pre-imputation GWAS analysis	119
3.10	Effect allele frequencies for the tagging SNPs used to reconstruct <i>ABO</i> subgroups	123
3.11	Fine mapping: 99% credible SNP set for the chromosome 9 association.....	127
3.12	MAGMA gene-set analysis	133
3.13	Putative GWAS associations: joint analysis group	135
4.1	Baseline group characteristics.....	144
4.2	Additional clinical phenotype data for the CTEPH and PE groups	146
4.3	ADAMTS13 and VWF antigen level pair-wise diagnostic group comparisons	149
4.4	Multivariable linear regression model of ADAMTS13 antigen levels	150
4.5	Multivariable linear regression model of VWF antigen levels	151
4.6	ADAMTS13 antigen level quartiles for CTEPH and healthy controls	153
4.7	Summary table for combined ADAMTS13 and VWF groups.....	154
4.8	Multivariable linear regression of ADAMTS13 plasma levels and interaction	

effects	155
4.9 Multivariable linear regression model of uncorrected (batch2) ADAMTS13 antigen levels.....	158
4.10 Multivariable linear regression model of VWF antigen levels and genetic ABO groups in CTEPH	174
4.11 Protein quantitative trait loci for ADAMTS13 antigen levels in CTEPH	175
4.12 Multivariable linear regression with the percentage of variance of ADAMTS13 antigen levels explained by SNPs and other characteristics	176
5.1 Missingness of variables from different recruitment centres in the GWAS minimal dataset following QC removals	187
5.2 VTE associated loci in the CTEPH GWAS	193
5.3 Warfarin metabolism associated loci in the CTEPH GWAS	197
5.4 Cox proportional hazards model of post-PEA survival in CTEPH	204

Table and figure legends are presented in boxes underneath the corresponding table or figure.

Abbreviations

Abbreviations and acronyms are defined at their first usage.

1000G	1000 Genomes (phase 3)
6mwd	6-minute walk distance
95% CI	95 percent confidence interval
Act	Activity
ADAMTS13	A disintegrin and metalloproteinase with a thrombospondin type 1 motif, member 13
Ag	Antigen
BF	Bayes factor
BPA	Balloon pulmonary angioplasty
BSA	Bovine serum albumin
CAD	Coronary artery disease
CADD	Combined Annotation Dependent Depletion
CBA	Collagen binding assay
CHR	Chromosome
CI	Cardiac index
CRP	C-reactive protein
CTED	Chronic thromboembolic disease
CTEPH	Chronic thromboembolic pulmonary hypertension
CTPA	Computed tomography pulmonary angiography
DHCA	Deep hypothermic circulatory arrest
DNA	Deoxyribonucleic acid
DVT	Deep vein thrombosis
EAF	Effect allele frequency
eCRF	Electronic case report form
ELISA	Enzyme-linked immunosorbent assays
eQTL	Expression quantitative trait loci
EV	Eigenvector
F11	Factor XI
FDR	False discovery rate
FGG	Fibrinogen Gamma

FRET	Fluorescence resonance energy transfer
FumaGWAS	Functional mapping and annotation of genome-wide association studies
GO	Gene ontology
GTE _x	Genotype-Tissue Expression
GWAS	Genome-wide association study
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
INFO	Information score
IPAH	Idiopathic pulmonary arterial hypertension
IQR	Interquartile range
Kb	Kilobase
LD	Linkage disequilibrium
MAF	Minor allele frequency
MAGMA	Multi-marker Analysis of GenoMic Annotation
Mb	Megabase
mPAP	Mean pulmonary artery pressure
NEA	Non-effect allele
NHP	Normal Human Plasma
NT-proBNP	N-terminal pro b-type natriuretic peptide
OPD	o-phenylenediamine dihydrochloride
OR	Odds ratio
PAH	Pulmonary arterial hypertension
PAWP	Pulmonary arterial wedge pressure
PBS	Phosphate buffered saline
PBST	Phosphate buffered saline plus Tween
PCA	Principal component analysis
PE	Pulmonary embolism
PEA	Pulmonary endarterectomy
PFT	Pulmonary function test
PH	Pulmonary hypertension
PP	Posterior probability
pQTL	Protein quantitative trait loci
PVR	Pulmonary vascular resistance

QC	Quality control
QoL	Quality of Life
QQ	Quantile-quantile
rhADAMTS13	Recombinant human ADAMTS13
RHC	Right heart catheterisation
rsID	reference SNP cluster ID
SD	Standard deviation
SE	Standard error
SNP	Single-nucleotide polymorphism
TTP	Thrombotic thrombocytopenic purpura
UTR	Untranslated region
VQ	Ventilation perfusion
VTE	Venous thromboembolism
VWF	Von Willebrand factor
WCC	White cell count
WHO	World Health Organisation
WTCCC	Wellcome Trust Case Control Consortium

1 Introduction

1.1 Literature search

A literature search was performed to ensure that there was a systematic approach to the evidence included in the introduction. A Healthcare Databases Advanced Search (HDAS) was conducted using the databases Medline and PubMed. The following search terms (including variations and acronyms) were used:

1. Chronic thromboembolic pulmonary hypertension AND
 - a. Aetiology OR pathobiology OR pathophysiology OR review OR update
 - b. Genetics OR genome-wide association study OR single nucleotide polymorphism
 - c. ADAMTS13 OR von Willebrand Factor OR haemostasis OR coagulation OR fibrinolysis
2. (Venous thromboembolism OR pulmonary embolism OR deep-vein thrombosis) AND (genome-wide association study OR single nucleotide polymorphism)
3. Genome-wide association study AND (review OR guideline)

References lists from articles were reviewed for additional citations. Literature and resources utilised during the PhD programme were also included where appropriate.

1.2 Chronic thromboembolic pulmonary hypertension

1.2.1 Overview

Pulmonary hypertension is defined as an increase in mean pulmonary arterial pressure (mPAP) ≥ 25 mmHg at rest.⁽¹⁾ Chronic thromboembolic pulmonary hypertension (CTEPH) is classified in group 4 pulmonary hypertension in the European Society of Cardiology / European Respiratory Society 2015 guidelines.⁽¹⁾ CTEPH can result from failure of thrombus resolution in the pulmonary arteries following acute pulmonary embolism (PE).⁽²⁾ Organisation and fibrosis of thrombotic material leads to obstruction of proximal pulmonary arteries and the subsequent development of a secondary small vessel (distal) vasculopathy, both of

which contribute to pulmonary hypertension and subsequent right heart failure.(3, 4) CTEPH is an infrequent but important complication of PE, a common disease affecting 1/1000, which increases with age.(5) In a meta-analysis, the pooled incidence of CTEPH was ~3% in survivors of PE.(6)

The diagnosis of CTEPH is based on international criteria and requires an mPAP \geq 25mmHg at right heart catheterisation (with a pulmonary arterial wedge pressure (PAWP) \leq 15mmHg) and specific radiological defects after at least 3 months of effective anticoagulation.(7) The most common imaging modalities to diagnose CTEPH are computed tomography pulmonary angiography (CTPA) and ventilation-perfusion (VQ) scans, with magnetic resonance angiography (MRA) and invasive pulmonary angiography performed less frequently ([Figure 1.1](#)).(8)

CTEPH represents the most severe long-term consequence of acute PE, but there are a range of more common manifestations within post-PE syndrome. Following PE, ~50% of patients will have functional limitations, 25-33% have persistence of thrombi and 10-30% have persistent or worsening right ventricular function / pulmonary artery pressures.(9) Also, within the post-PE spectrum is chronic thromboembolic disease (CTED), which is characterised by persistent pulmonary arterial thromboembolic occlusions without pulmonary hypertension (mPAP $<$ 25mmHg) in symptomatic patients.(10)

Risk factors for CTEPH include previous venous thromboembolism (VTE), with preceding PE and deep vein thrombosis (DVT) occurring in 75% and 50% of patients respectively in European cohorts.(11) The risk of developing CTEPH following VTE is higher with recurrent VTE, unprovoked VTE, larger PEs and right heart dysfunction at the time of the acute PE.(12-14) Additional risk factors include: non-O blood groups, splenectomy, antiphospholipid antibodies / lupus anticoagulant, malignancy, thyroid replacement therapy and a putative association with ventriculo-atrial shunts / infected pacemakers.(12-14)

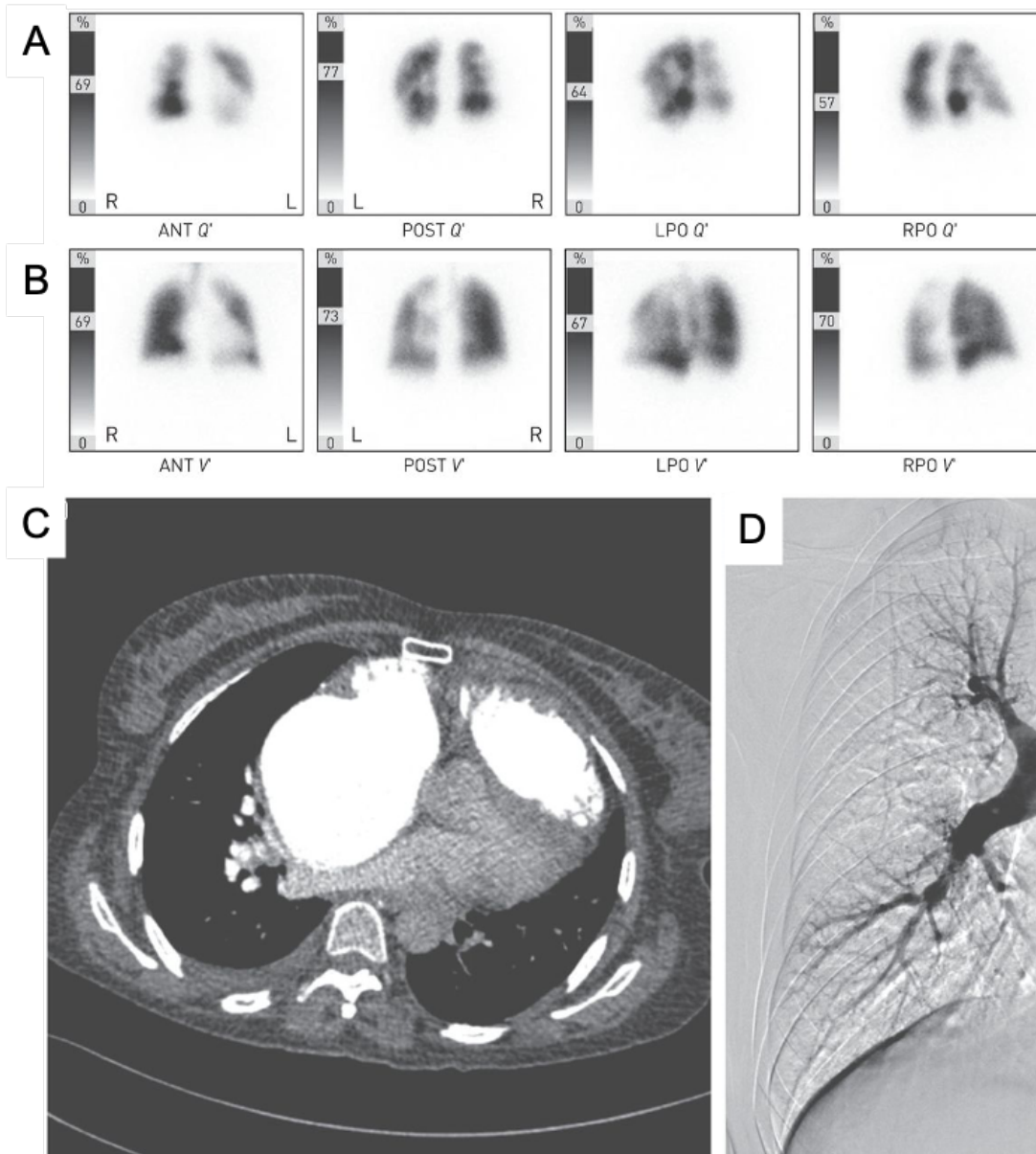


Figure 1.1 Imaging modalities for diagnosing CTEPH

Different imaging modalities in a patient with CTEPH deemed to be operable. Perfusion (**A**) and ventilation (**B**) (VQ) scans showing typical mismatched perfusion defects. **C** Computed tomography pulmonary angiogram (CTPA) displaying marked right ventricular dilation and a paucity of pulmonary arteries in the lower lobes. **D** Invasive pulmonary angiogram showing web and stenotic lesions in the pulmonary arteries of the right lung. ANT (anterior), POST (posterior), LPO (left posterior oblique), RPO (right posterior oblique). Figure reproduced with permission of the © ERS 2020, from (19).

1.2.1.1 CTEPH management

Management of CTEPH consists of lifelong anticoagulation.(1) Traditionally, vitamin-K antagonists (e.g. warfarin) have been used for anticoagulating CTEPH patients and more recently direct oral anticoagulants (DOACs).(16) The main side effect of long-term anticoagulation treatment is major and clinically relevant non-major bleeding however, in CTEPH the major bleeding rate is low (<0.7%/person-year).(16) There is no evidence that additional treatments including thrombolytic medications that result in blood clot dissolution, can prevent CTEPH following acute pulmonary embolism.(17, 18)

Suitable CTEPH patients should have a multi-disciplinary evaluation of surgical operability. Eligible patients are offered pulmonary endarterectomy (PEA), which involves removing obstructive thromboembolic material in surgically accessible pulmonary arteries.(7) The procedure requires a midline sternotomy followed by cardiac bypass and deep hypothermic cardiac arrest (DHCA) to minimise bleeding in the surgical field.(19) The obstructive thromboembolic material together with the intima and superficial medial layer of the pulmonary artery can be removed to the level of the subsegmental pulmonary arteries if required ([Figure 1.2](#)).(19) PEA surgery is the optimal treatment in eligible patients, offering the best chance of symptomatic and prognostic improvement.(7) Patients have marked improvements in haemodynamics, World Health Organisation (WHO) functional class, 6-minute walk distance (6mwd), patient reported outcomes including quality of life, and survival following PEA.(17, 20)

Percutaneous balloon pulmonary angioplasty is a CTEPH treatment that involves dilating the stenotic pulmonary arterial segments with a balloon catheter. It is a treatment modality that can be considered for selected CTEPH patients ineligible for PEA due to distal disease or alternatively, in those with persistent or recurrent pulmonary hypertension (PH) following PEA.(17, 21-23)

Patients with surgically inaccessible disease, disproportionate distal predominant disease or with co-morbidities precluding PEA can be treated with licensed pulmonary artery vasodilator medication and emerging therapies.(24, 25)



Figure 1.2 Pulmonary endarterectomy specimen

Pulmonary endarterectomy specimen removed from right and left pulmonary arteries. Figure reproduced with permission of the © Elsevier, from (26)

1.2.2 CTEPH pathobiology

1.2.2.1 CTEPH histology

CTEPH occurs due to proximal occlusion of pulmonary arteries by organised fibrotic clots and a micro-vasculopathy in smaller pulmonary vessels.(3, 27) Resected PEA specimens show remodelling of thrombi at various stages with different degrees of inflammation and cellularity, which is distinct from the fresh clot found in PE specimens.(2, 28) In a series of 54 CTEPH patients undergoing PEA, the specimens were comprised of: collagen (100%), elastin (67%), haemosiderin (56%), atherosclerosis (32%) and calcification (15%), with inflammation present in 53%.(29) Plexogenic lesions (a dynamic network of vascular channels) are a hallmark of idiopathic pulmonary arterial hypertension (IPAH) and have been identified in some studies of CTEPH, suggesting a degree of histopathological overlap with other types of pulmonary hypertension.(3, 30) A more recent study utilising PEA specimens, CTEPH transplant tissue and a porcine model identified the importance of post-capillary remodelling (venopathy) and anastomosis between the pulmonary and systemic circulation in CTEPH.(31)

1.2.2.2 CTEPH pathobiology: introduction

There are a number of unanswered questions related to the pathobiology of CTEPH. Potential mechanisms and pathways that are involved include: excess thrombus formation, failure of fibrinolysis, endothelial dysfunction, failure of neovascularisation, inflammation, endothelial dysfunction and others including right ventricular adaptation and genetics.(2, 27, 32)

1.2.2.3 Increased thrombus

A thrombus (blood clot) is the final product of blood coagulation. Coagulation involves a complex biological cascade and is a component of haemostasis, the process of stopping bleeding. Haemostasis comprises vascular spasm, platelet activation and coagulation. Following damage to the endothelium of a blood vessel, platelets are activated and recruited to the site of injury to form a platelet plug (described further in **Section 1.6**) (primary haemostasis).(33) Injury to the endothelium leads to the release of tissue factor and a cascade of coagulation serine proteases that ultimately results in the formation of a cross-linked fibrin clot (secondary haemostasis) (**Figure 1.3**).(33)

The observation that the risk of CTEPH is increased with larger and recurrent PEs is consistent with increased thrombus formation being important in CTEPH pathobiology.(12, 14) Plasmatic factors associated with thrombus formation are increased in CTEPH including antiphospholipid antibodies / lupus anticoagulant, von Willebrand Factor (VWF) / Factor VIII, tissue-type plasminogen activator (t-PA) and type 1 plasminogen activator inhibitor (PAI-1) (**Figure 1.4**).(34-36) Increased thrombus formation alone is unlikely to be the sole contributing factor to CTEPH pathobiology and other studies have suggested that traditional thrombophilic risk factors and antiphospholipid antibodies are not over-represented in CTEPH (see **Section 1.3.1**).(34, 37) Furthermore, ~25% of patients with CTEPH will not have a preceding (known) VTE and thrombolysis treatment of acute pulmonary embolism that decreases the thrombus burden has no effect on the development of CTEPH.(11, 18)

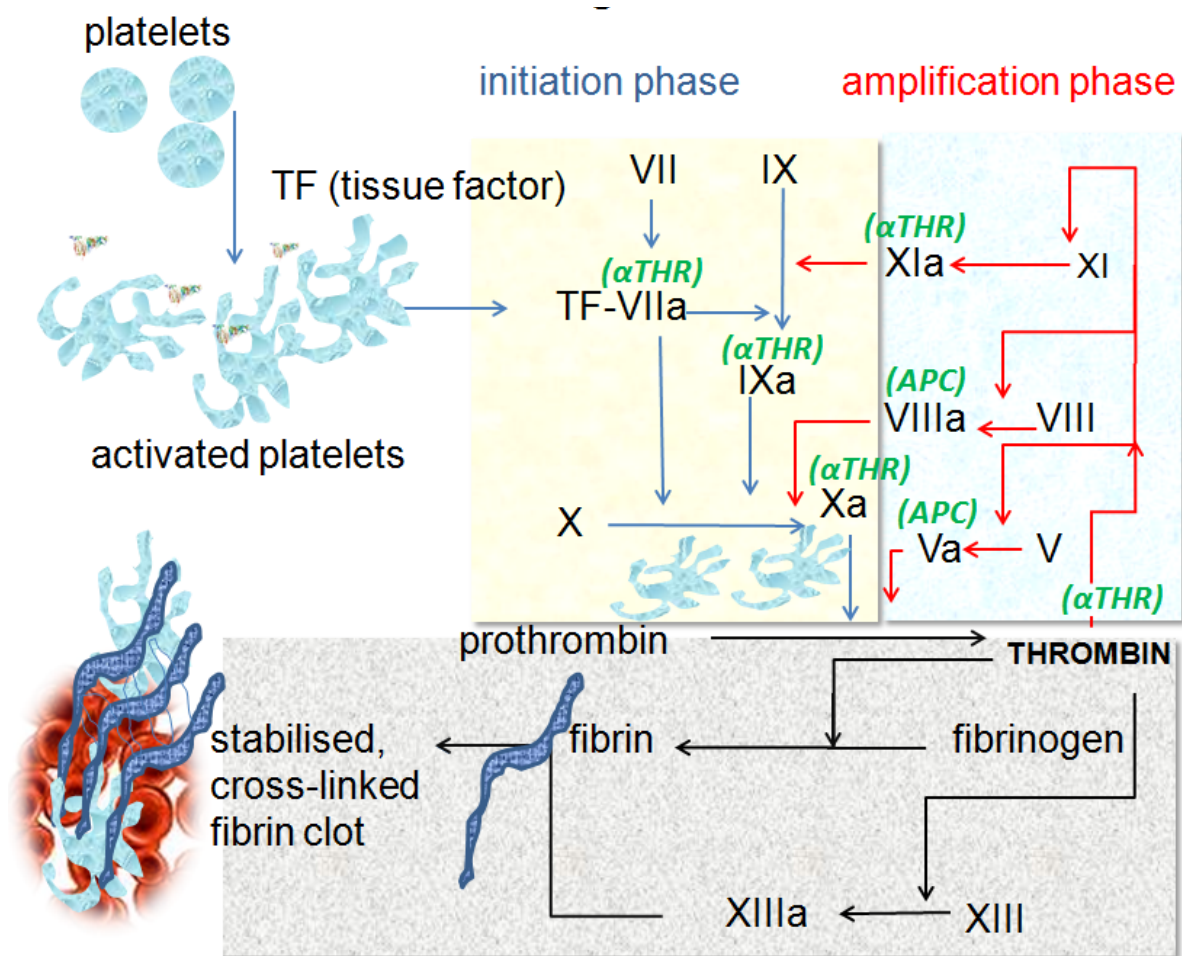


Figure 1.3 Blood coagulation overview

Roman numerals (i.e. VII) denote clotting factors and a lowercase “a” denotes the active form. The coagulation cascade has two components, the initiation phase (extrinsic pathway) and the amplification phase (intrinsic pathway). APC (activated protein C), TF (tissue factor), αTHR (antithrombin).

Figure reproduced under creative commons license, from (38).

1.2.2.4 Fibrinolysis

If fibrinolytic mechanisms are involved in CTEPH pathobiology this may result in reduced thrombus dissolution (**Figure 1.4**). Fibrin from CTEPH patients is partially resistant to plasma mediated lysis suggesting impaired fibrinolysis.(39) Five fibrinogen variants have been described in CTEPH that result in differences in the molecular structure of fibrin.(40) However, fibrin resistance is not specific to CTEPH and occurs in other forms of pulmonary hypertension and to a lesser extent PE

suggesting an epiphenomenon.(41) There may be differences in fibrinogen genetic polymorphisms which account for fibrin resistance in CTEPH that is discussed further in [Section 1.3.2](#).(42) The fibrinolysis inhibitor thrombin-activatable fibrinolysis inhibitor (TAFI) is increased in CTEPH and plasma levels correlate with resistance to clot lysis, suggesting an additional fibrinolysis pathway that may be affected in CTEPH.(43)

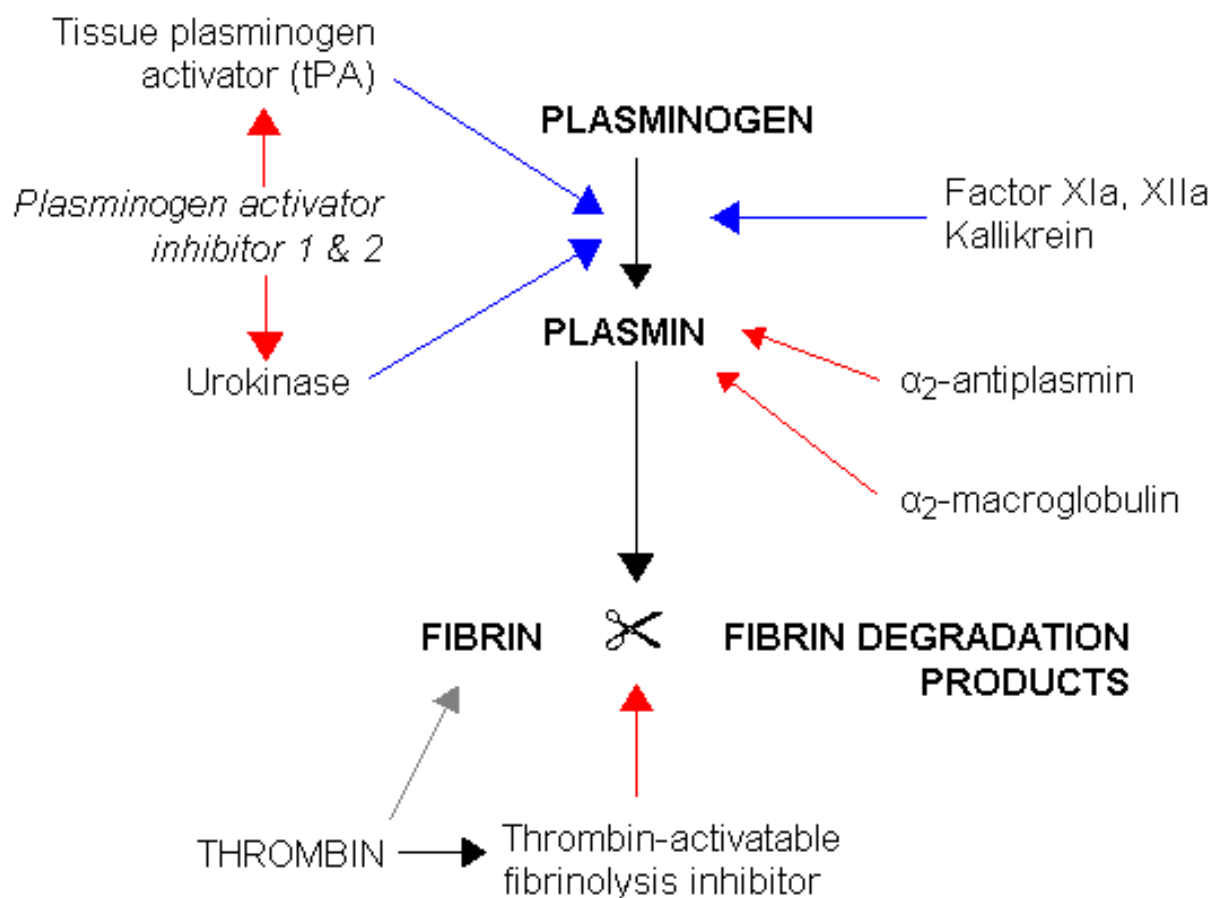


Figure 1.4 The fibrinolysis pathway

A simplified diagram of the fibrinolysis pathway. Fibrinolysis is the process of breaking down a fibrin blood clot formed during coagulation. Blue arrows indicate stimulation and red inhibition.

Figure reproduced under creative commons license, from (44)

1.2.2.5 Neoangiogenesis

Restoration of pulmonary artery patency following thrombus obstruction is important following an acute pulmonary embolus and abnormalities in neoangiogenesis may be important in CTEPH pathobiology. PEA specimens have reduced vascular structures and lower expression of vessel-specific genes.(45) Impaired neoangiogenesis is associated with increased mortality and persistent PH following PEA.(46) Furthermore, factors that inhibit neoangiogenesis (e.g. platelet factor 4) are increased in PEA specimens, which may result in altered calcium homeostasis and endothelial dysfunction.(47) These findings suggest neoangiogenesis is defective in CTEPH, although it is unclear whether this is causative.

1.2.2.6 Inflammation and endothelial dysfunction

Inflammation is important in other forms of pulmonary hypertension (e.g. PAH) and may be involved in CTEPH pathobiology.(48) Some risk factors for CTEPH are also associated with inflammation including cancer, inflammatory bowel disease and infected indwelling lines.(12, 49) Plasma C-reactive protein (CRP) is increased in CTEPH and may contribute to mechanisms involved in fibrotic vascular remodelling and endothelial dysfunction.(50, 51) Inflammatory cells (macrophages, T-lymphocytes and neutrophils) have been found to accumulate in PEA specimens and correlate with raised plasma CRP.(46) Inflammatory mediators are also upregulated in PEA specimens including interleukin-6 (IL-6), monocyte chemoattractant protein-1 (MCP-1), interferon- γ -induced protein-10 (IP-10) and macrophage inflammatory protein (MIP)-1 α . A number of circulating inflammatory markers are elevated in CTEPH including MCP-1, IL-6, IL-8, IP-10, MIP-1 α and matrix metalloproteinase 9.(46, 52, 53)

1.2.2.7 Problems investigating CTEPH pathobiology

Elucidating the pathobiology of CTEPH has been challenging for several reasons. Firstly, the haemostasis and fibrinolytic pathways have been implicated but are difficult to study in anticoagulated CTEPH patients. Secondly, whilst risk factors for progression from acute PE to CTEPH have been described, identifying causative pathological mechanisms in the post-PE disease spectrum would require large well-characterised prospective PE cohorts.(54) Lastly, animal models do not appropriately recapitulate failure of thrombus dissolution or the chronic changes that occur in

CTEPH.(55) Therefore, an alternative approach to investigating CTEPH pathobiology and circumventing these issues is to perform exploratory genetic studies.

1.3 Genetics of CTEPH

1.3.1 Traditional thrombophilia risk factors

A thrombophilia is an abnormality of blood coagulation that increases the risk of thrombosis and can either be acquired or inherited.(56) The traditional heritable thrombophilias can be divided into loss of function mutations (antithrombin, protein C and protein S deficiencies) and gain of function mutations (prothrombin gene and factor V Leiden mutations).(57)

1.3.1.1 Heritable thrombophilia overview

The factor V Leiden (FVL) mutation is the most common heritable thrombophilia in the UK with the heterozygous form affecting 3-7% of Caucasian European populations.(56) It is caused by a missense variant (rs6025) in the clotting factor V (*F5*) gene.(58) The FVL mutation results in activated factor V (FVa) being resistant to the proteolytic effects of activated protein C (APC resistance) leading to increased FVa levels and increased generation of thrombin ([Figure 1.5](#)).(56)

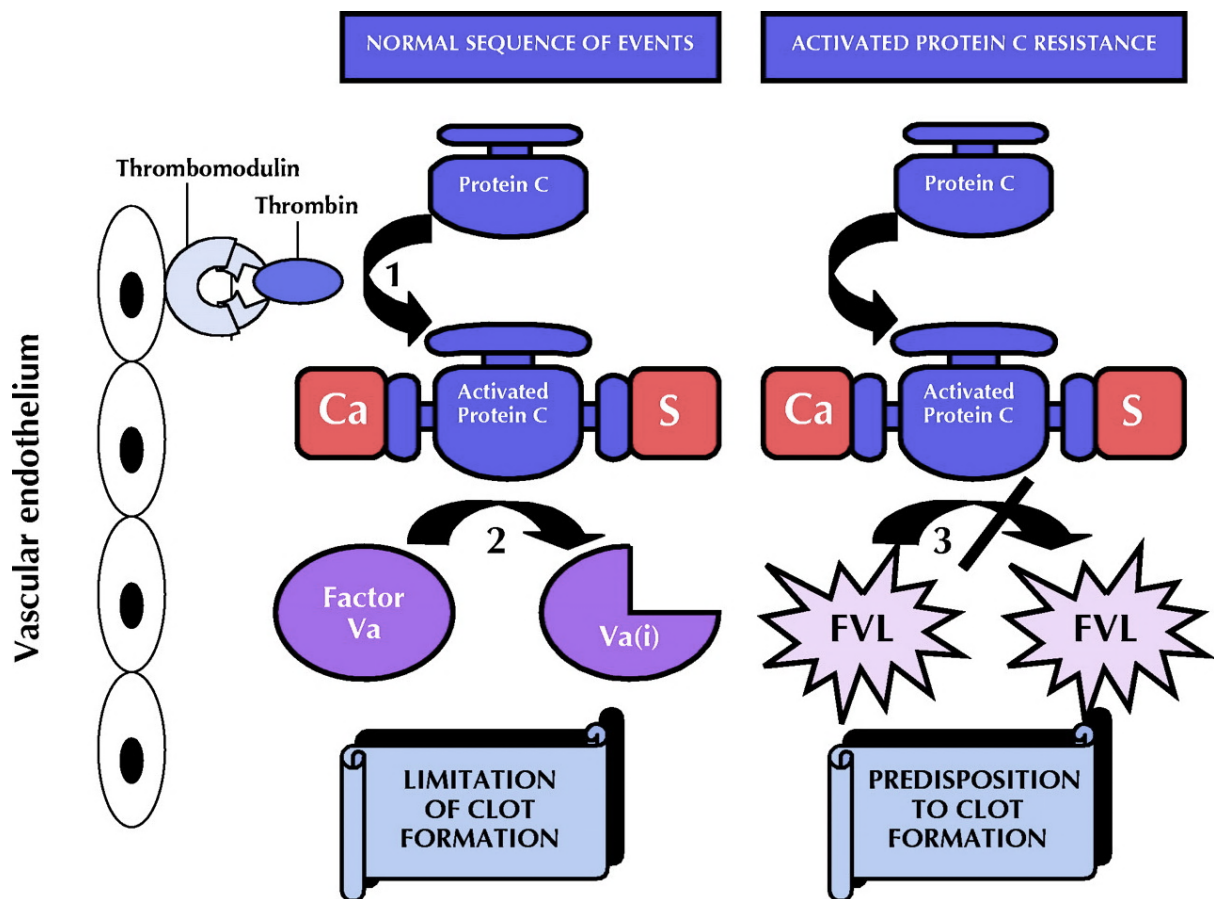


Figure 1.5 Activated protein C and the factor V Leiden mutation

Activated protein C (in the presence of protein S (S) and calcium (Ca)) will usually inactivate factor Va. When the FVL mutation is present, FVa is resistant to the proteolytic effects of activated protein C predisposing to clot formation

Ca (ionised calcium), S (protein S).

Figure reproduced with permission of the © CMAJ group, from (59)

The prothrombin gene mutation is the second most common heritable thrombophilia in the UK affecting 1-2% of Caucasian Europeans.(56) It is caused by a 3' untranslated region (UTR) variant (guanine to adenine; rs1799963) in the prothrombin (*F2*) gene at nucleotide position 20210 (termed the prothrombin G20210A mutation).(60) This results in an increase in prothrombin (factor II) levels and a hypercoagulable state (**Figure 1.6**).

Protein C and S are both vitamin-K dependent glycoproteins that have anticoagulant properties.(58) Inherited protein C and S deficiencies affect approximately 0.3% and

0.1% of Caucasian European populations respectively.(56) Deficiencies in either protein C or S lead to impaired inactivation of factors Va and VIIIa resulting in increased thrombus formation (**Figure 1.6**).⁽⁶¹⁾ Most mutations causing protein C / S deficiencies are heterozygous with homozygous being very rare and often fatal. There have been over 160 different protein C gene (*PROC*) and ~200 protein S gene (*PROS1*) mutations described that exhibit an autosomal dominant pattern of inheritance.^(62, 63)

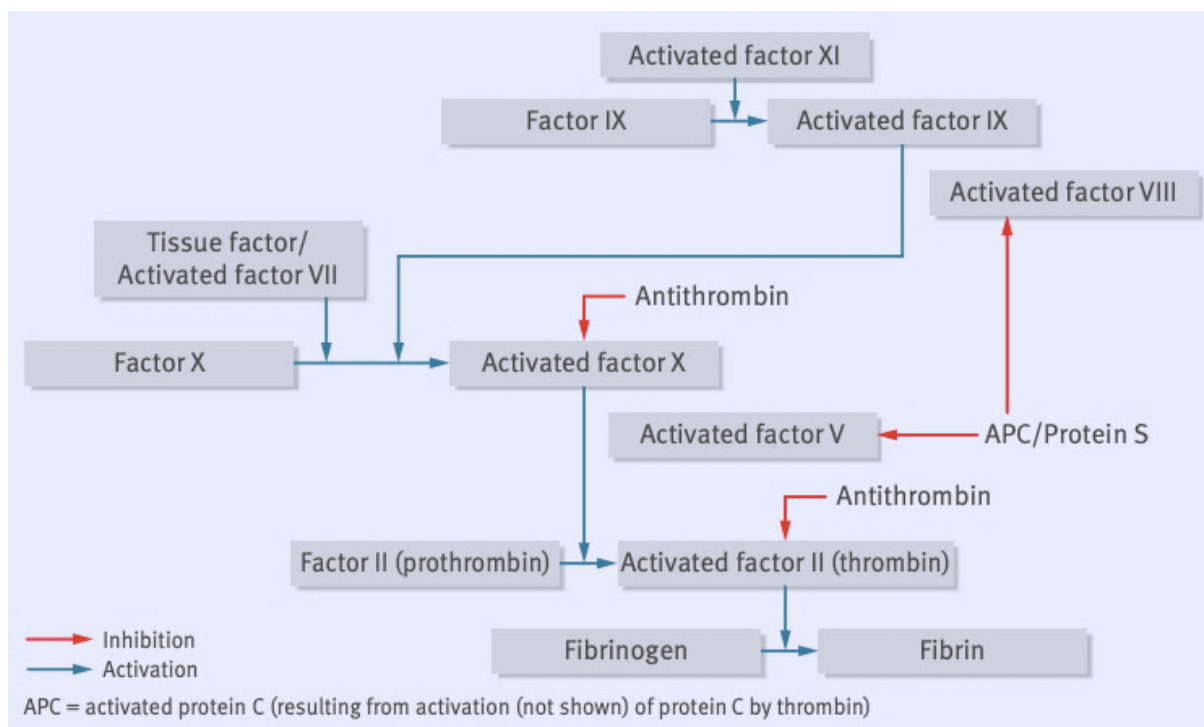


Figure 1.6 Activation and inhibition in the coagulation system

Overview of the blood coagulation cascade with additional details related to inhibition (red arrows) and activation (blue arrows)

Figure reproduced with permission of the © BMJ Publishing Group Ltd, from (56).

Antithrombin is the main natural inhibitor of blood coagulation (**Figure 1.6**).⁽⁶⁴⁾ Antithrombin deficiency affects approximately 0.02% of Caucasian European populations.⁽⁵⁶⁾ Deficiency in antithrombin leads to reduced inhibition of factors Xa and thrombin, thereby creating a hypercoagulable state (**Figure 1.6**). The majority of mutations that lead to antithrombin deficiency are heterozygous with over 130

described in the *SERPINC1* gene and exhibiting an autosomal dominant pattern of inheritance.(64)

The heritable thrombophilias are summarised in [Table 1.1](#) and the risk of VTE associated with heritable thrombophilias are described in [Section 1.4.1](#).

Thrombophilia	Prevalence (%)	Type of mutation / inheritance	Mechanism of thrombosis
FVL, heterozygous	4	missense	FVa resistant to proteolysis by activated protein C
Prothrombin (G20210A) gene mutation	1.5	missense	Increased prothrombin (factor II) levels
Protein C deficiency	0.3	>160 mutations Autosomal dominant	Impaired inactivation of factors Va and VIIIa
Protein S deficiency	0.1	~200 mutations Autosomal dominant	Impaired inactivation of factors Va and VIIIa
Antithrombin deficiency	0.02	>130 mutations Autosomal dominant	Reduced inhibition of factors Xa and thrombin

Table 1.1 Heritable thrombophilias

Summary of prevalence (in a Caucasian European population), type of mutation and inheritance pattern and the consequence by which the thrombotic risk is increased.

FVL (factor V Leiden).

Adapted from (56)

1.3.1.2 Heritable thrombophilias and CTEPH

Genetic associations have been inferred from observational studies that have identified CTEPH risk factors. Thrombotic risk factors (not including ABO) occurred in 27.7% of 426 CTEPH patients in an international prospective register.(11) Of the

heritable thrombophilias, 9.6% and 8.9% had protein S and C deficiencies, 7.7% had the factor V Leiden (FVL) mutation, 3.5% had the prothrombin gene (G20210A) mutation and 0.7% had antithrombin III deficiency (some patients had more than one thrombophilia).(11)

Early studies found no difference in the prevalence (see below) of FVL mutation between CTEPH and PAH or between CTEPH and healthy controls.(34, 65) A study by Wolf *et al*, investigated the heritable thrombophilic risk factors between 46 patients with CTEPH and 100 healthy controls and found no statistical difference.(34) However, the FVL mutation occurred in 6.5% of CTEPH patients and 3% of healthy controls, and antitrypsin III deficiency, protein S deficiency and protein C deficiency did not occur at all in CTEPH or healthy controls. The study conclusions were limited by a lack of power to detect group differences, which is compounded by the rarity of some of the heritable thrombophilias (see [Section 1.3.1.1](#)). In a study by Wong *et al*, 29% of CTEPH patients (n=45 total) were heterozygous for the FVL mutation compared with 8% in a group with other forms of pulmonary hypertension (n=200 total).(37) Study participants were all Caucasian, which is important as the FVL mutation occurs in ~5% of Caucasian populations but is rare in other ethnic groups.(37) There was no difference in antithrombin III deficiency, protein C / S deficiency and prothrombin gene mutations between CTEPH and control groups.(37) There was a nominally significant difference in the frequency of the FVL mutation in another study by Suntharalingam *et al* of 214 CTEPH patients and 200 healthy controls (FVL: 3.6% versus 1.5%).(66) There was no difference in antithrombin III deficiency, factor XIII, plasminogen activator inhibitor and tissue plasminogen activator polymorphisms between CTEPH and the healthy control groups.(66) Both the Wong *et al*, and Suntharalingam *et al* studies were underpowered and only examined a limited number of genetic variants.

1.3.2 ABO and fibrinogen genes

ABO groups are another heritable thrombotic risk factor. Non-O blood groups are more prevalent in CTEPH than the general population occurring in 76% and 54% respectively.(67, 68) The frequency of non-O blood groups in CTEPH and VTE appears similar across studies.(67) However, there has not been a direct study

comparison in the same population to determine if ABO groups are different in resolved PE *versus* CTEPH.

A genetic variant (rs6050) in the *FGA* gene encoding Fibrinogen A α chain protein is more common in CTEPH than healthy controls. (42, 66) A study of 101 patients with CTEPH and 102 with pulmonary embolism in a Han Chinese population reported an over-representation of this *FGA* polymorphism in CTEPH compared to PE. This study was limited by the inability to adjust for population stratification (a potential difference in the allele frequencies of subpopulations; discussed further in [Section 1.5](#)) and small sample size for genetic associations. Furthermore, rs6050 has been associated with VTE in genetic studies and the *FGA-FGB-FGG* gene region is associated with VTE in genome-wide association studies ([Section 1.5](#)). (69-71) The allele (a variant of a gene at a given location) frequencies of the rs6050 *FGA* genetic variant between CTEPH and VTE have not been investigated in an adequately powered study.

1.3.3 Genetic variants associated with pulmonary arterial hypertension in CTEPH

Genetic variants associated with pulmonary arterial hypertension (PAH) have been investigated in CTEPH and are described below ([Section 1.3.3.2](#)) together with the background to PAH ([Section 1.3.3.1](#)) for context and comparison with the genetics of CTEPH.

1.3.3.1 Pulmonary arterial hypertension overview

PAH is classified as group 1 (pre-capillary) pulmonary hypertension in the European Society of Cardiology / European Respiratory Society 2015 guidelines.(1) The pathobiological processes in PAH involve endothelial dysfunction, pulmonary arterial smooth muscle proliferation and accumulation of fibroblasts, myofibroblasts and inflammatory cells in the pulmonary arterial wall.(72) The net effect is pulmonary vascular remodelling and vessel obstruction resulting in pulmonary hypertension and eventually right heart failure.(72) It is diagnosed according to right heart catheter haemodynamic criteria (mPAP > 20mmHg, a pulmonary vascular resistance (PVR) \geq 3 Wood units and a PAWP \leq 15mmHg) in the absence of other causes of pre-capillary PH (e.g. left heart disease, chronic lung disease, CTEPH).(1, 73) PAH has an

incidence of 2.0 - 7.6 cases per million adults per year with a female preponderance affecting them four times more often.(74) There are a number of sub-types of PAH including idiopathic PAH (IPAH), heritable PAH (HPAH), drug / toxin induced PAH and PAH associated with other diseases (e.g. connective tissue disease or congenital heart disease) (**Table 1.2**). (73) Current licensed medical treatments are pulmonary artery vasodilating medications that aim to improve the vasoconstriction, but do not alter the underlying disease pathobiological mechanisms. PAH is an incurable disease with a median survival of 6 years, which varies depending on the particular type of PAH.(74)

1 PAH
1.1 Idiopathic PAH
1.2 Heritable PAH
1.3 Drug- and toxin-induced PAH
1.4 PAH associated with:
1.4.1 Connective tissue disease
1.4.2 HIV infection
1.4.3 Portal hypertension
1.4.4 Congenital heart disease
1.4.5 Schistosomiasis
1.5 PAH long-term responders to calcium channel blockers
1.6 PAH with overt features of venous/capillaries (PVOD/PCH) involvement
1.7 Persistent PH of the newborn syndrome

Table 1.2 Clinical classification of PAH (group 1 PH)

PAH (pulmonary arterial hypertension), PVOD (pulmonary veno-occlusive disease), PCH (pulmonary capillary haemangiomatosis)

Table from (73)

HPAH describes patients with an identified germline mutation or a family history of PAH (not associated with other diseases).(75) In 70–80% of families with PAH and 10–20% of IPAH cases the cause is a mutation in the gene encoding the bone morphogenetic protein receptor type II (*BMPR2*). (76) There are over 400

heterozygous germline *BMPR2* mutations that have been identified to date.(75) *BMPR2* is a receptor for the transforming growth factor- β (TGF- β) superfamily and is expressed on the surface of a wide range of cells particularly the pulmonary vascular endothelium. Mutations in *BMPR2* cause abnormal signalling that can adversely impact endothelial barrier function, DNA repair, cell proliferation, inflammation, metabolism and mitochondrial function.(72) A number of other rare gene variants have been identified for PAH and are summarised in **Figure 1.7**. There have also been common genetic variant associations with PAH that are described further in **Section 1.3.4**.

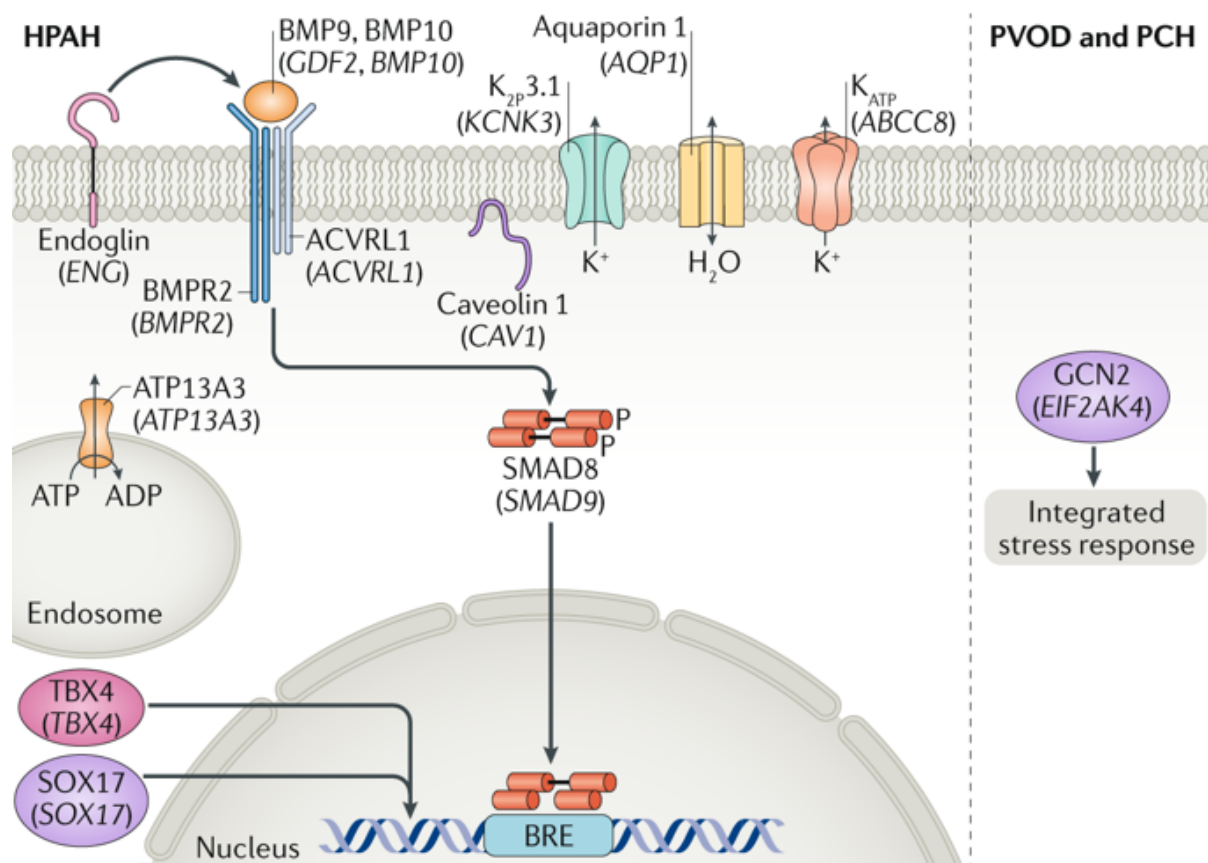


Figure 1.7 HPAH pathways and gene variants

Rare gene variants that have been associated with HPAH are summarised in the figure (in addition SMAD1 and SMAD4; not shown). EIF2AK4 mutations are associated with PVOD (pulmonary veno-occlusive disease) / PCH (pulmonary capillary haemangiomatosis).

ABCC8 (ATP binding cassette subfamily C member 8), ACVRL1 (activin A receptor like type 1), ADP (Adenosine diphosphate), AQP1 (Aquaporin 1), ATP (Adenosine triphosphate), ATP13A3 (ATPase 13A3), BMP (bone morphogenetic protein), BRE (BMP-responsive element), CAV1 (Caveolin 1), EIF2AK4 (Eukaryotic translation initiation factor 2 alpha kinase 4), ENG (Endoglin), GCN2 (general control nonderepressible 2), GDF2 (growth differentiation factor 2), KCNK3 (potassium two pore domain channel subfamily K member 3), SMAD (SMAD family member), SOX17 (SRY-box transcription factor 17), TBX4 (T-box transcription factor 4).

Figure reproduced with permission of the © Springer Nature, from (75)

1.3.3.2 Pulmonary arterial hypertension genetic variants and CTEPH

Genetic variants that occur in other types of pulmonary hypertension have been investigated in CTEPH. A study of 49 patients with CTEPH and 17 with PE in a Han Chinese population reported a higher frequency of variants in genes associated with PAH including *BMPR2*, *ACVRL1*, *ENG*, *SMAD9*, *CAV1*, *KCNK3*, and *CBLN2*.(77) However, this candidate gene approach was potentially confounded by the small sample size, incomplete assessment of variant deleteriousness and inability to adjust for population stratification. Larger studies in Caucasian CTEPH cohorts have not identified mutations in the bone morphogenetic protein type II receptor (*BMPR2*) gene that occur in heritable and idiopathic pulmonary arterial hypertension.(78, 79) Prostacyclin is an important regulator of vascular tone and proliferation, and variants in genes related to prostacyclin have been implicated in PAH.(80) A study in 90 CTEPH patients and 144 healthy controls in a Japanese cohort investigated a variable-number tandem repeat (VNTR) polymorphism in the 5'-upstream promoter region of the *PGIS* gene that has also been associated with PAH.(80, 81) However, this specific VNTR *PGIS* gene promotor variant was not associated with CTEPH.(81) Importantly, some candidate genes (e.g. *PGIS*) putatively associated with PAH and used in studies of CTEPH genetic associations, have not been replicated in larger studies of PAH using robust methodology.(82) In summary, genetic variants associated with PAH do not convincingly occur in CTEPH in the limited studies performed.

1.3.3.3 Other CTEPH genetic associations

There are apparent differences in CTEPH between Japanese and Caucasian populations with a female preponderance, less preceding DVT and a lower overall incidence of CTEPH in Japanese patients.(83) Consequently, studies have investigated the human leukocyte antigen (HLA) region in Japanese CTEPH cohorts and potential associations in the *HLA-DPB1* (DPB1*0202 allele) and *NFKBIL1* (IKBLP*03 allele) genes have been identified, which require validation.(84, 85) The HLA class II histocompatibility antigen, DP(W2) beta chain (HLA-DPB1) protein is expressed on antigen presenting cells and plays a key role in the immune system, whereas the role of the NF-kappa-B inhibitor-like protein (NFKBIL1) protein is unclear but may also be involved in immune function as it is located in the major histocompatibility complex class I region.(86, 87) In a separate Japanese study of 97 patients with CTEPH, variants in the *ACE* gene that expresses the protein angiotensin-converting enzyme, which has a role in vascular remodelling and endothelial dysfunction, were investigated.(88) There was no difference in genotype frequencies between CTEPH and healthy controls.(81, 88) A small, unvalidated whole exome sequencing study in 30 patients with CTEPH in a Chinese population identified one patient with a *MUC6* missense variant (rs201234174), a gene encoding a glycoprotein expressed by epithelial tissues.(89)

1.3.4 CTEPH genetic architecture

There are limited observational case reports of presumed “familial” CTEPH however, genotyping was not performed to establish shared genetic mutations.(90-92) A recent study of 66 prevalent CTEPH cases used genealogy data (without genotypes) to infer genetic relationships.(90-92) They found that whilst a classical Mendelian inheritance was not observed in CTEPH, there was an excess of familial clustering, which implies shared risk factors that may be genetic.(90-92) Interestingly, the relative risk of VTE was increased in the 1st degree relatives of CTEPH patients, which could suggest an excess of heritable thrombophilias.(90-92) Whether the familial clustering was due to shared genetic risk factors for VTE, CTEPH or shared environmental risk factors remains an unanswered question.

Whilst the genetic architecture of CTEPH has not been fully elucidated in robust studies, a Mendelian inheritance pattern is not clearly observed. As CTEPH

predominantly occurs following VTE, its genetic architecture may mirror VTEs, which is primarily a complex polygenic disease ([Section 1.4](#)). In complex, polygenic diseases, common genetic variants exert small effect sizes on disease risk compared with highly penetrant Mendelian diseases whereby rare mutations exert large effects on disease risk ([Figure 1.8](#)).

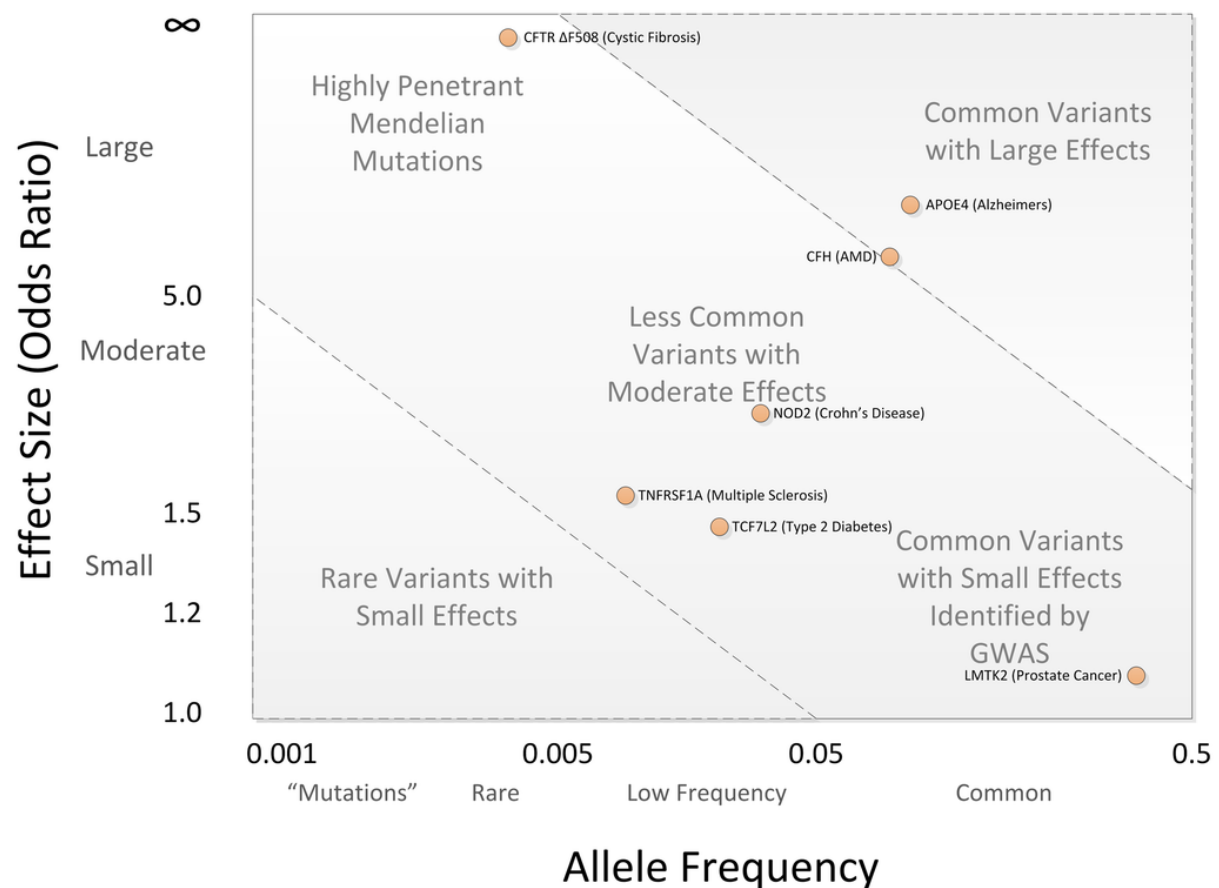


Figure 1.8 The disease effect size of alleles with varying frequencies

The disease effect size (Odds ratio) of an allele is plotted against the frequency of the allele. In highly penetrant Mendelian diseases (top left of the figure) rare mutations exert large effects on the disease risk, whereas in most polygenic diseases investigated with genome-wide associations studies, common variants exert small effects on disease risk (bottom right of figure).

Figure reproduced with permission of the Copyright: © 2012 Bush, Moore, from (93).

CTEPH can be juxtaposed with the genetic architecture of PAH, which is a Mendelian disorder with an autosomal dominant inheritance and incomplete penetrance, whereby not all individuals with the genetic mutation (e.g. in *BMPR2*) exhibit the PAH phenotype.(76) The incomplete penetrance suggests there are additional causative factors that may include additional genetic, epigenetic or environmental contributors.(76) Recent studies have identified common genetic variant associations with PAH in the *HLA-DPA1/DPB1* region (rs2856830) and near the *SOX17* (SRY-Box Transcription Factor 17) gene locus (rs10103692 and rs10103692).(94) The *HLA-DPA1/DPB1* gene region is within the major histocompatibility class II region and involved in the immune system, whereas the *SOX17* gene has a pivotal role in angiogenic processes including the development of the lung microvasculature.(75)

1.3.5 CTEPH epigenetics

Differences in gene expression and epigenetic mechanisms that do not involve changes in the DNA sequence have been implicated in CTEPH pathobiology.(95, 96) These include changes in microRNAs, DNA methylation and transcription factor expression.(97-99) These epigenetic studies are also limited by small sample sizes, lack of replication cohorts and establishing the functional consequences of the potential epigenetic differences.(95)

1.3.6 Genetics of CTEPH: summary

Overall, genetic studies in CTEPH have been small, underpowered and unvalidated. Familial clustering suggests shared risk factors, which may be genetic. A Mendelian pattern of inheritance does not occur in CTEPH, and it is more likely to follow a complex disease paradigm with common variants exerting small to moderate effects. This would be consistent with CTEPH being a severe consequence of VTE, itself a polygenic disease, which is described in [Section 1.4](#). Non-O blood groups are over-represented in CTEPH, but other heritable thrombophilias and PAH associated gene variants have not been conclusively associated. To fully understand CTEPH genetic associations, the genetics of VTE need to be considered.

1.4 Genetics of venous thromboembolism

1.4.1 Traditional thrombophilia risk factors

There are a number of risk factors for venous thromboembolism including environmental (e.g. cancer, surgery, trauma, immobilisation, pregnancy, age, body mass index, medications) and genetic.(100) VTE is highly heritable with family-based studies estimating the variation in VTE that is attributable to genetics to be ~60%.(101) This is consistent with the incidence of VTE varying with ethnicity, being higher in Caucasians and Africans compared with Asians.(102)

Genetic risk factors for VTE include the traditional heritable thrombophilias ([Section 1.3.1](#)). Loss of function mutations in the anticoagulant genes (antithrombin, protein C and protein S deficiencies) increase the risk of VTE 5-20 fold but they are rare, being present in <1% of the population.(103) Gain of function (heterozygous) mutations in procoagulant genes (prothrombin (G20210A), and factor V Leiden mutation) have a more moderate impact on VTE risk of 2-5 fold.(57) In a meta-analysis of 126,525 patients with VTE from 173 case-control studies, the significant risk associations were: FVL mutation (OR 9.45 95% CI 6.72-13.30) and prothrombin gene mutation (OR 3.17; 95% CI 2.19-3.46).(104) The VTE risk from the FVL mutation is lower in subsequent genetic studies (see [Section 1.4.3](#)), and varies from 5 fold in heterozygotes to 50 fold in homozygotes.(103) The risk of VTE with different hereditary thrombophilias is summarised in [Table 1.3](#). Fibrinogen variants (*FGG* gene) are also associated with VTE (OR 2.4; 95% CI 1.5-3.9) and result in a decrease in plasma levels of fibrinogen gamma.(105) A combination of more than one hereditary thrombophilias has a synergistic effect on VTE risk. In a pooled analysis of 2310 VTE cases and 3204 healthy controls, patients with heterozygous FVL and prothrombin gene mutations had a 20 fold increased risk of VTE (95% CI 11-36).(106)

Heritable thrombophilias can further increase the risk of VTE when combined with other pro-thrombotic risk factors including pregnancy, and oestrogen containing medications (e.g. the oral contraceptive pill (OCP) and hormonal replacement therapy (HRT)). In pregnancy, a number of clotting factors (I, II, VII, VIII, IX and XII) are increased, and fibrinolysis is inhibited, which increases the risk of VTE.(59) In a case-control of pregnant women 44% with VTE had the FVL mutation (heterozygous)

compared to 8% without VTE (relative risk (RR) 9; 95% CI 5-17).(107) In the same study, 17% of pregnant women with VTE had the prothrombin (G20210A) gene mutation compared to 1% without (RR 15; 95% CI 4-53).(107) The risk of VTE is also increased in pre-menopausal women with the FVL mutation that are taking oestrogen containing OCPs (RR 35; 95% CI 8-154) and post-menopausal women with the FVL mutation taking oestrogen containing HRT (RR 7; 95% CI 3-14).(108, 109) However, whilst the relative risks are synergistically increased by the combination of heritable thrombophilias and pregnancy or oestrogen containing medications, the absolute risk remains low, particularly in the latter.(110)

Thrombophilia	VTE relative risk
FVL, heterozygous	3-5
FVL, homozygous	10-50
Prothrombin (G20210A) gene mutation	2-4
Antithrombin deficiency	10-20
Protein C deficiency	5-15
Protein S deficiency	10

Table 1.3 Risk of VTE with different hereditary thrombophilias

The relative risk of VTE with different heritable thrombophilias. Heterozygous is the carriage of a mutation of only one allele, whereas homozygous is of two alleles. FVL (Factor V Leiden). Adapted from (56, 103, 110)

Studies on VTE risk factors including genetic associations, will often consider DVT and PE together rather than separately. This is partly due to their co-occurrence, with (predominantly asymptomatic) PE occurring in 40-50% of proximal DVT, and (predominantly proximal and asymptomatic) DVT occurring in 70% of patients with PE.(111) Therefore, to define isolated DVT (DVT without PE) or isolated PE (PE without DVT) for comparative studies would require imaging of both the lower limbs (e.g. doppler ultrasonography) and lungs (e.g. CTPA), which is not routinely performed in clinical practice. Nevertheless, differential risk factors between DVT and PE have

been observed. The FVL mutation occurs more frequently in DVT than PE, which is termed the factor V Leiden paradox.(112) A meta-analysis identified significantly more patients with DVT had the FVL mutation compared to those with isolated PE (OR 2.39; 95% CI 2.08-2.75).(113) The mechanism by which the FVL mutation exerts its increased risk on DVT is unclear.(112)

Common variant genetic associations in venous thromboembolism that have been discovered by genome-wide associations studies are described in [Section 1.5.5](#).

1.4.2 Venous thromboembolism and ABO

Non-O blood group is the most important population attributable genetic risk factor for VTE, as it occurs more commonly than other heritable thrombophilias, despite conferring a lower individual risk (OR 2.09; 95% CI 1.83-2.38).(114) In a US study, non-O blood group occurred in 64.2% of VTE cases (n=492) compared with 52.5% of controls (n=1008).(115) Furthermore, the combination of non-O blood group and a thrombophilia has a supra-additive effect on VTE risk.(115) Interestingly, non-O blood groups are more common in recurrent and idiopathic VTE, which are also risk factors for progression from acute PE to CTEPH.(116, 117)

The ABO groups vary by the ABH(O) antigens (oligosaccharide residues) and are found on red blood cells, platelets and VWF, a protein involved in haemostasis and described in [Section 1.6](#).(118) The ABO gene encodes glycosyltransferase enzymes that transfer specific monosaccharides to the H precursor antigen.(119) Individuals with blood groups A and B have terminal sugar moieties consisting of N-acetylgalactosamine and d-galactose, respectively, whereas those with blood group O lack transferase enzyme activity and therefore only express the H antigen.(120) Whilst the exact mechanism(s) linking ABO antigen group to thrombotic risk has not been defined, it may be mediated by VWF levels, which are 25% lower in O group individuals.(121)

1.5 Genome-wide association studies

Performing a genome-wide association study in CTEPH is an opportunity to investigate genetic associations that may inform disease pathobiology. GWAS is a

method of examining genetic variants across the genome of individuals to identify associations with a disease or trait. This section will provide an overview of the background, methods, statistics and challenges of GWAS.

1.5.1 GWAS background

GWAS has been transformative in identifying genetic variants associated with multiple common traits.(122) Common (polygenic) traits are caused by a combination of many genetic and environmental factors in contrast to Mendelian single-gene disorders.(123)

In contrast to linkage studies in highly penetrant genetic conditions that identify causative variants, GWASs have found many genetic variants with relatively small effect sizes that predispose to disease.(124) Family-based linkage studies utilise family pedigrees with the disease of interest together with genetic markers to identify associated genes based on the concept that chromosomal regions co-segregate in families.(93) However, linkage analysis is limited by low power for complex traits that are associated with multiple genes and the challenge in narrowing down causative variants from large co-segregated chromosomal regions. (93, 124) GWAS predominantly investigate the genetic variation in single nucleotide polymorphisms (SNPs), which are single base pair changes in the DNA sequence.(93) There are over 80 million SNPs in the human genome with an individual's genome varying from a reference genome by 4-5 million SNPs. (125) Common variants have traditionally been defined as those with a minor allele frequency (MAF; for a biallelic SNP the frequency of the second most common allele) $> 5\%$, low-frequency variants are those with a MAF $1-5\%$ and rare variants have a MAF $<1\%$.(126) More recently, common variants have been defined as those with a MAF $\geq 1\%$ and rare variants having a MAF $<1\%$, although as allele frequencies are continuous the cut-offs are arbitrary.(122) As individual GWAS associations are predominately common SNPs with modest effect sizes, they are neither necessary nor sufficient to cause disease.(124) The concept that common diseases have a different genetic architecture to rare Mendelian and are associated with common variants is termed the common-variant common disease hypothesis.(127)

The first GWAS was performed in 2005 and identified a common genetic variant in the complement factor H gene in age-related macular degeneration (AMD).(128) There have subsequently been over 5000 GWAS performed for over 70,000 variant-trait associations ([Figure 1.9](#)).(129) Whilst the first GWAS included only 96 cases with AMD, modern GWAS from the largest consortia have included hundreds of thousands of cases and a continued increase in case numbers is likely.(128, 130) GWAS is superior to candidate gene approaches (for example a study of genes associated with thrombosis and fibrinolysis), which are limited by widespread false-positive associations and the inability to evaluate unknown (*a priori*) genetic variants.(131) With a hypothesis free approach, GWAS has identified multiple genes that were not previously implicated in disease pathogenesis and variants in genomic regions containing no genes (intergenic).(132) The transition from candidate gene studies to GWAS for the investigation of complex diseases was possible due to the reduced cost of SNP genotyping over the last two decades.(124)

GWAS has many different applications in addition to identifying SNP-trait associations that can inform disease pathobiology. The proliferation of international consortia and biobanks (i.e. DeCODE, UK Biobank) have enabled larger GWAS to be performed that have identified an increasing number of common variant associations.(133, 134) These risk variants can be combined into polygenic risk scores together with traditional environmental risk factors to predict an individual's disease susceptibility, a tenet of personalised medicine.(135) GWAS can inform clinical treatments by identifying genetic variants that are associated with efficacy, drug metabolism and side effects.(93) Another benefit of large biobank cohorts is they contain a number of diseases and traits that enables genetic variants of interest to be studied across different phenotypes in phenome-wide association studies (PheWAS).(136) PheWAS has shown that many individual genetic variants are associated with multiple traits, which is termed pleiotropy.(137) Risk variants identified by GWAS can be used as genetic instruments in Mendelian randomisation studies, the process of investigating causal relationships between potentially modifiable risk factors (e.g. CRP) and health outcomes (e.g. coronary artery disease).(137-139) An overview of the different genetic applications for GWAS is summarised in [Table 1.4](#).

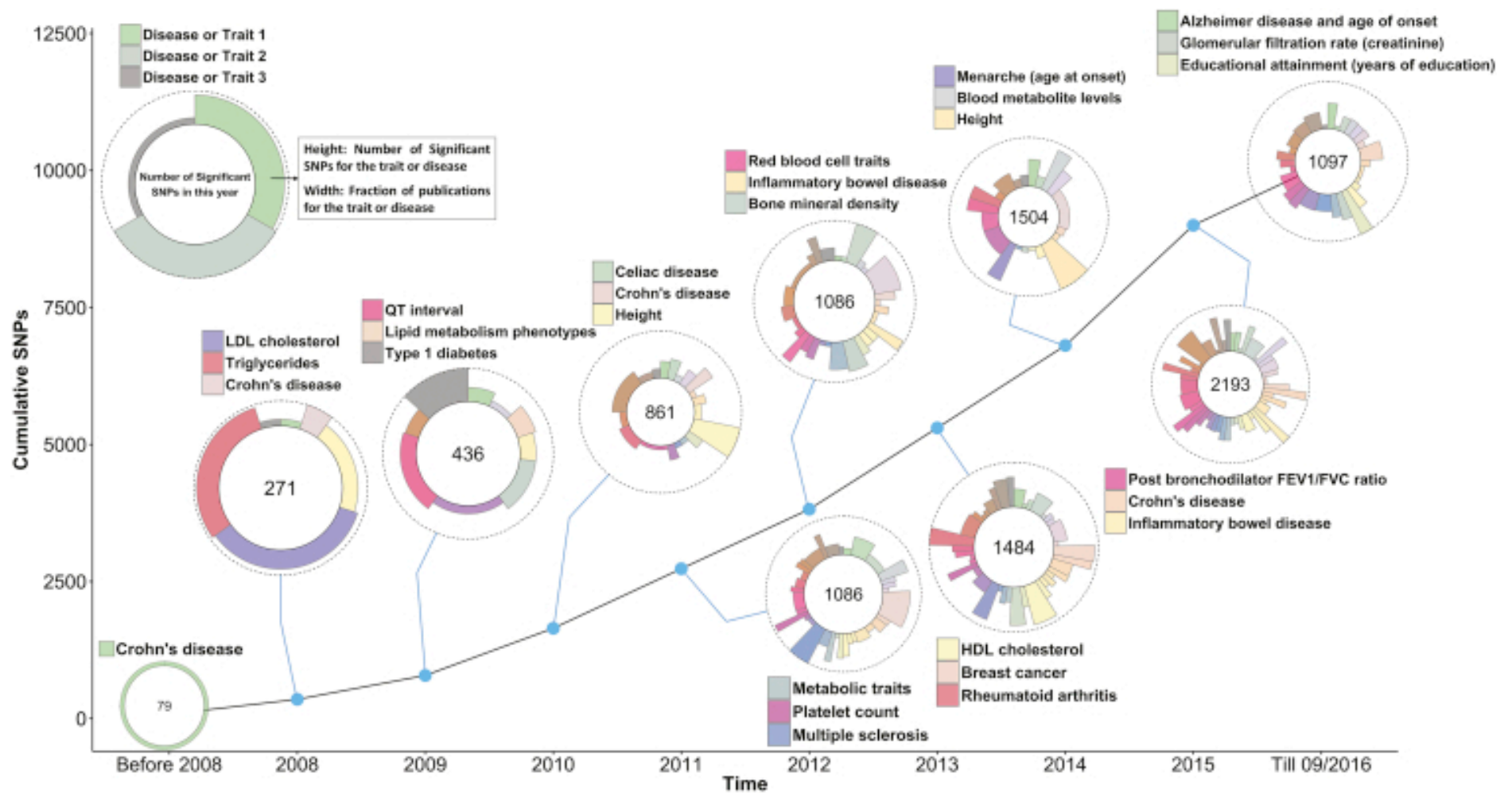


Figure 1.9 GWAS SNP-trait associations have increased over time

Data for the GWAS associations was from the GWAS Catalog.(140) SNPs were included if they had a $p < 5 \times 10^{-8}$ and were not in linkage disequilibrium (LD; see [Section 1.5.2](#)) ($r^2 < 0.5$) with another associated SNP (if they were in LD, the SNP with the lowest p -value was retained). Only the top three traits and diseases with the largest number of SNP associations for each year have been included to improve visualisation.

Figure from (122)

Analysis	Purpose
GWAS	Detecting trait-SNP associations
Estimation of SNP heritability	Evaluation of genetic architecture
Population differences in allele frequencies	Reconstructing human population history and detecting evidence of natural selection
Polygenic risk scores	Detecting pleiotropy; validating GWAS discoveries
Mendelian randomization	Testing causal relationships
Trait GWAS with -omics GWAS	Fine-mapping; detecting target genes and functional consequence of genetic variant

Table 1.4 GWAS genetic applications

SNP heritability involves estimating the proportion of genetic variation captured by common SNPs. Pleiotropy is the concept that individual genetic variants are associated with multiple traits. Fine-mapping is discussed further in [Section 1.5.3](#). In addition, GWASs have a role in estimating genetic correlation to detect pleiotropy, detect copy number variant (CNV) – trait associations, and quantify the genome architecture by assessing linkage disequilibrium ([Section 1.5.4](#))

Table adapted from (122)

1.5.2 GWAS methods

Since the understanding that candidate gene studies produced many false positive associations, GWAS methodology has been developed to be systematic and robust with a focus on quality control and reducing bias. The stages include sample preparation, microarray clustering and genotyping of SNPs, sample and SNP quality control, statistical analysis and validation/replication of results. Additional steps include genetic imputation and meta-analysis, which are discussed in [Section 1.5.3](#).

Genomic DNA is extracted from individual samples (e.g. whole blood) and quantitatively assessed to ensure sufficient DNA for genotyping. SNPs are genotyped using high-throughput microarrays, whereby wells are coated with an allele-specific

oligonucleotide probe (short DNA sequence) that binds with sample DNA fragments and is then labelled with fluorescent dyes to produce an intensity signal.⁽¹⁴¹⁾ The signals are detected and then converted to genotypes (**Figure 1.10**). SNP microarrays vary in the number of SNPs they genotype, which is usually 200,000 to over 2 million.⁽¹²²⁾ Whilst GWASs still predominantly use SNP microarrays, there is increased utilization of SNPs from whole genome sequencing for association testing and this is likely to increase in the future.⁽¹²²⁾

Following genotype calling, comprehensive quality control is undertaken for individual samples and SNPs followed by any necessary exclusions to reduce potential confounding.⁽¹⁴²⁾ The quality control steps are further explained in detail in **Chapter 2**. In a case control study design, the different allele frequencies can then be compared between disease cases and healthy controls or a disease comparator group. This should then be followed by a confirmation of associated SNP-traits in a separate validation cohort.⁽¹⁴³⁾ An overview of the GWAS process is shown in **Figure 1.11**.

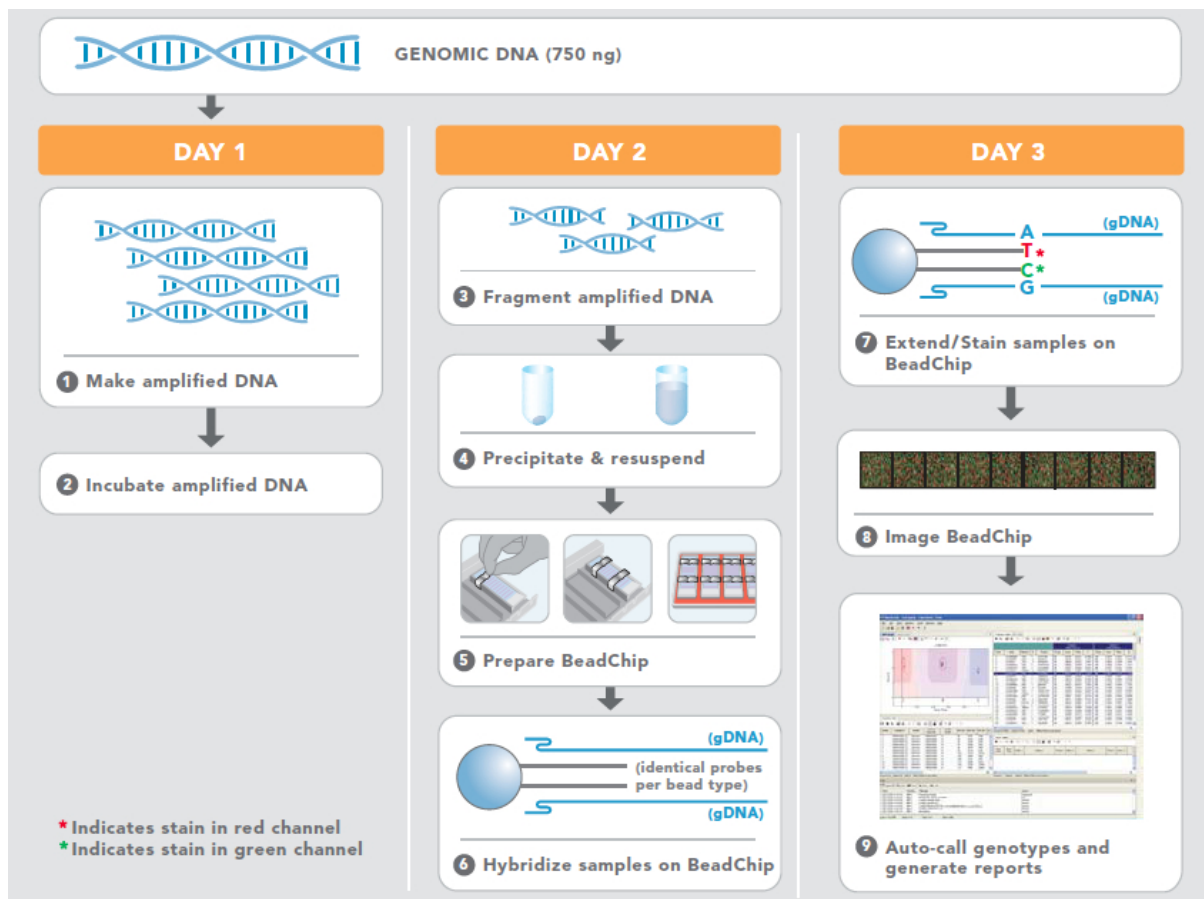


Figure 1.10 SNP microarray overview

Genomic DNA (gDNA) is amplified (left panel), followed by DNA fragmentation and direct hybridisation of sample DNA with microarray (in this example “BeadChip”) bound oligonucleotides (short DNA sequences) (middle panel) and finally fluorescent staining to produce an intensity signal (red or green) that is computationally analysed to provide genotype calls (right panel).

Figure reproduced with permission of the © Illumina from Illumina Infinium 2 assay schema, (<https://dnatech.genomecenter.ucdavis.edu/infinium-assay/>; accessed 22/1/20)

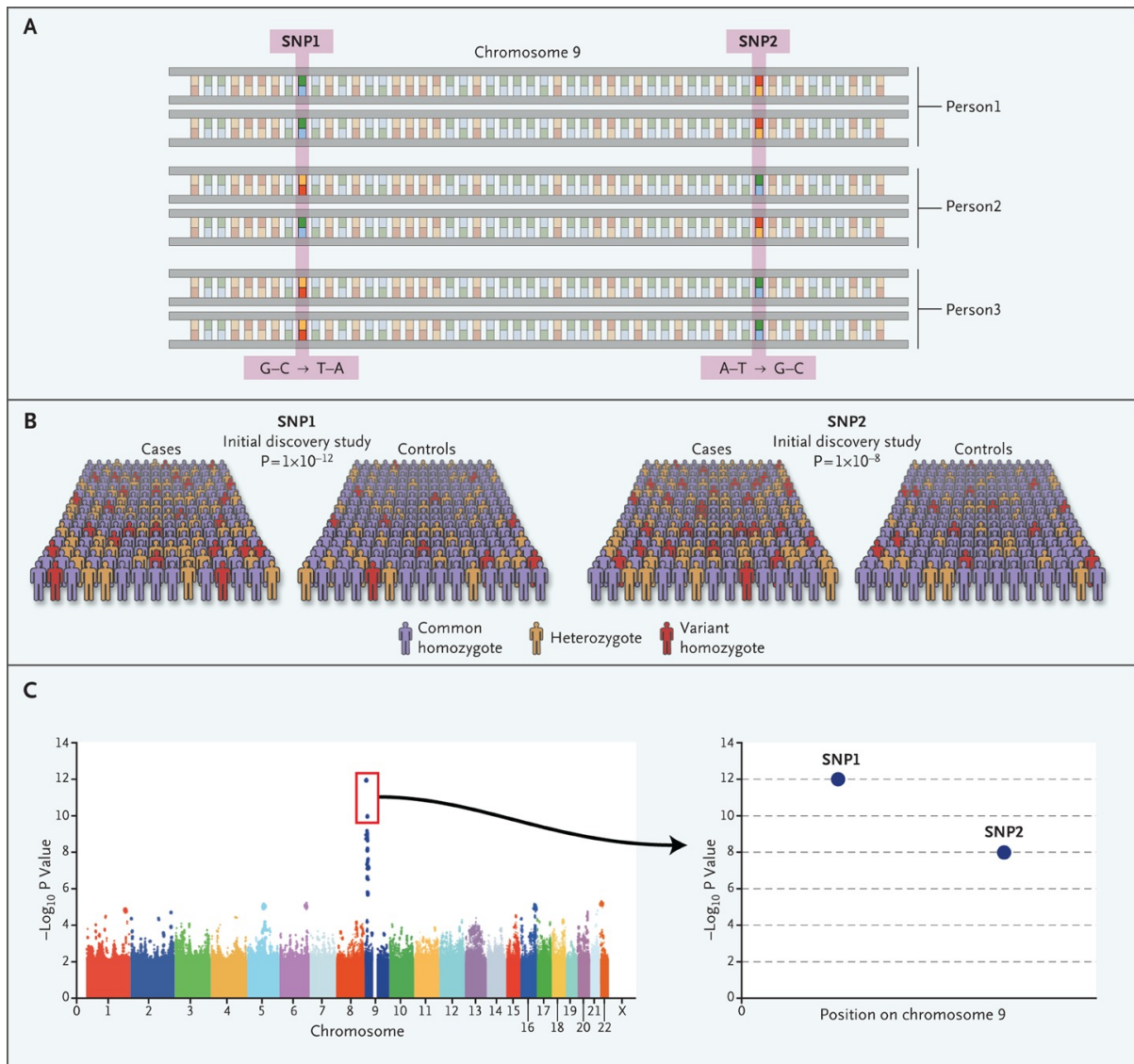


Figure 1.11 GWAS overview

A SNPs across the genome are microarrayed for each individual (small segment of chromosome 9 displayed)

B In a case-control GWAS, each SNP (SNP1, SNP2, ..., SNP_x) is compared between cases and controls. For a biallelic hypothetical SNP there would be two alleles (A/a) and 3 potential genotypes (AA (common homozygote), Aa (heterozygote), aa (variant homozygote)). The statistical comparison results in a *p*-value for each SNP based on the allele frequencies between cases and controls.

C Following quality control steps, the SNP associations (*p*-values) are visualised on the y-axis of a Manhattan plot, with SNP genomic position plotted on the x-axis

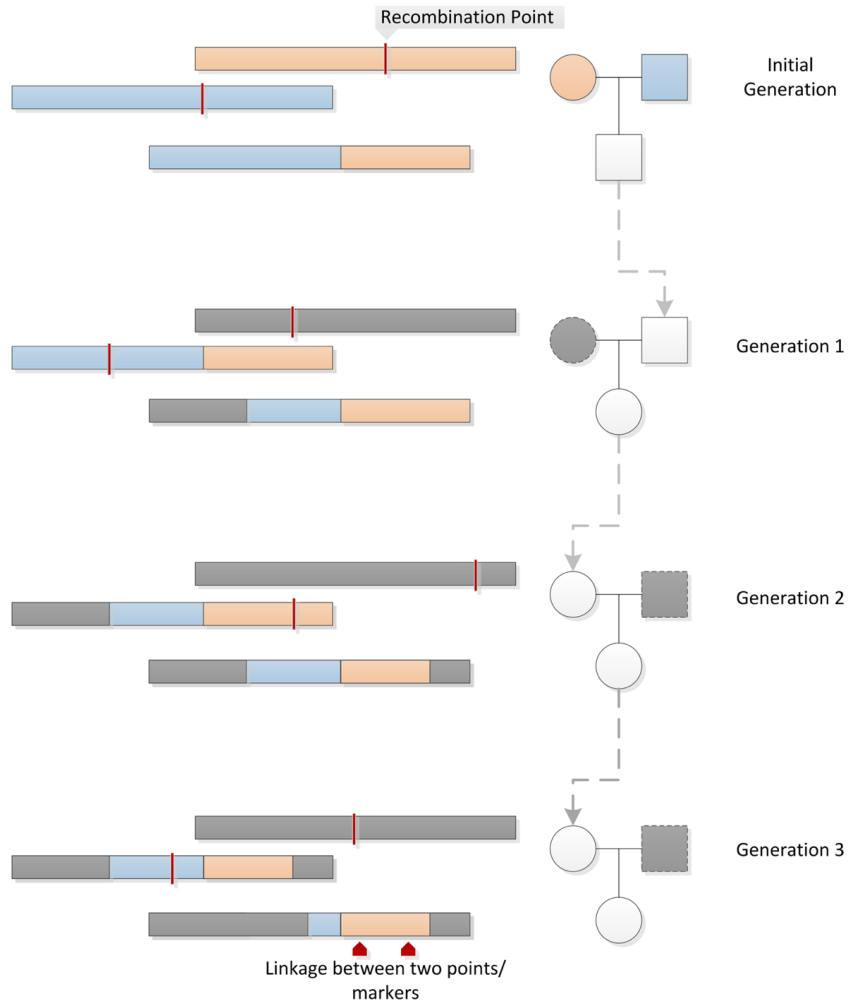
Figure reproduced with permission of the © Massachusetts Medical Society, from (123)

1.5.3 GWAS association testing and statistics

GWASs are able to capture variation across the whole genome without directly measuring all SNPs due to linkage disequilibrium (LD), which is the non-random association of alleles at two or more loci ([Figure 1.12](#)).⁽¹⁴⁴⁾ Alleles at flanking loci tend to be inherited together, with specific combinations termed haplotypes.⁽¹⁴⁵⁾ Whilst a single genotyped (tag) SNP is unlikely to have direct functional (causal) effects, it may be in LD with common variants that do, thus acting as a proxy marker.⁽¹⁴⁶⁾ GWASs rely on LD to infer SNP associations without genotyping them all individually. As genetic variation and associations are affected by ancestry, samples with different ancestry could introduce confounding from genotype differences between (and within) cases and controls that are related to population (ethnicity) differences rather than SNP-trait associations.^(147, 148) This population structure can result in a difference in allele frequencies due to ancestry rather than SNP-trait associations and needs to be accounted for in the analysis.

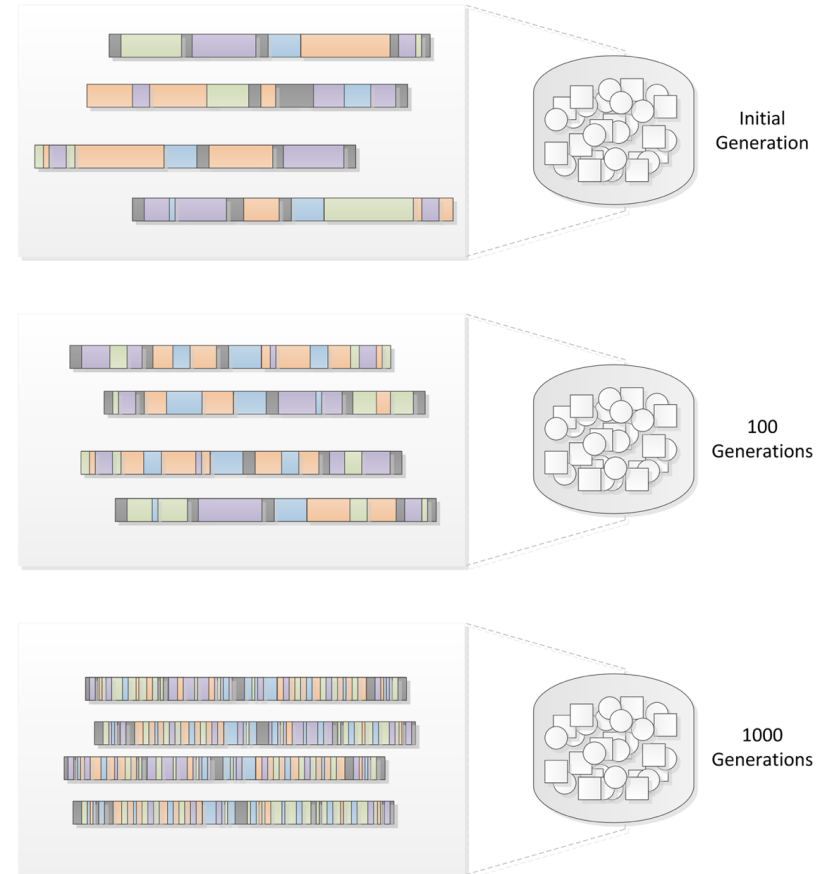
After quality control steps and any appropriate exclusions, association testing is performed. In a case-control study, allele frequencies can be statistically compared between cases and controls using a simple chi-squared test. However, there are usually other potential confounders (including population structure) that require adjusting for and therefore, multivariable logistic regression (for binary case/control) or linear regression (for continuous traits) is performed. GWAS case-control association testing results are often presented as an odds ratio for each SNP. For a hypothetical biallelic SNP (A/a) the allelic OR would represent the odds of disease in an individual with allele A compared with the odds of disease in an individual carrying allele a.⁽¹⁴⁵⁾

Linkage Within A Family



Linkage Disequilibrium Within A Population

Decay of Linkage over successive generations



Population moves from Linkage Disequilibrium to Linkage Equilibrium over time

Figure 1.12 Linkage and linkage disequilibrium

Within families, when a pair of genetic markers remains linked due to co-segregation of chromosomal regions rather than being separated by recombination events (red lines) then linkage has occurred (left panel). In contrast, in populations where there are a number of recombination events over time, genetic markers move from linkage disequilibrium to linkage equilibrium (right panel). Linkage disequilibrium is the non-random association of alleles at two or more loci.

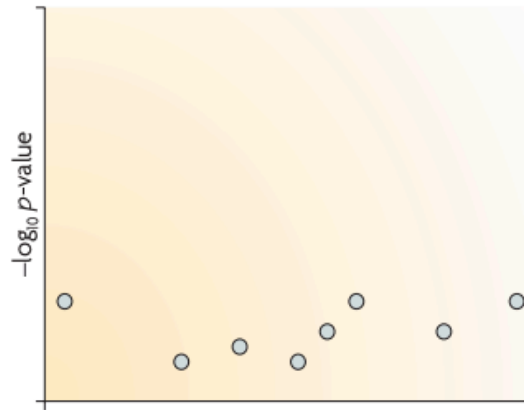
Figure from (93)

In a GWAS, the number of genotyped SNPs ranges from hundreds of thousands to millions and a statistical test is performed on each, which creates a problem of multiple testing and false positive associations.(132) This is addressed in GWAS by adjusting p -values for multiple testing or setting a more stringent threshold for p -value significance. This is often performed using the Bonferroni correction method, whereby the p -value (alpha) is divided by the number of SNPs (i.e. $0.05 / 1,000,000 = p < 5 \times 10^{-8}$). A p -value $< 5 \times 10^{-8}$ has become the standard to denote genome-wide significance.(143) GWAS association testing results are often visualised using Manhattan plots as shown in [Figure 1.14](#). In addition to a case-control GWAS, sub-phenotypes and continuous variables within diseases and traits can be explored using GWAS methodology.

The SNP coverage across the genome can be increased with genetic imputation, which is the statistical inference of unobserved genotypes that have not been directly measured by microarrays ([Figure 1.13](#)). (149) The measured alleles are compared with reference haplotypes (i.e. HapMap) and a statistical probability is estimated for the unmeasured genetic variants.(149) This allows more genetic variants to be tested for phenotype associations.

The increased utilisation of GWAS has led to results from different studies for the same diseases or traits. These results can be pooled together in a GWAS meta-analysis using the individual level genotype data or the summary statistics generated from association testing.(150) This increases sample size and the power to detect genetic variants with smaller effects on disease risk.

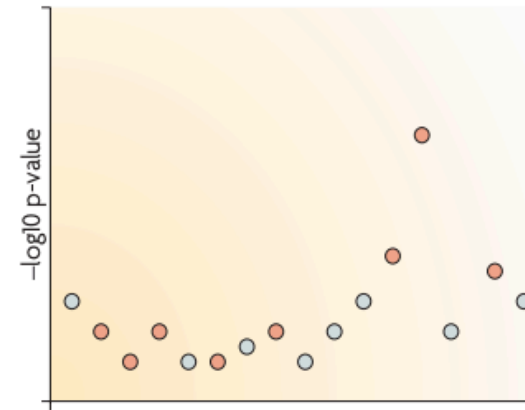
b Testing association at typed SNPs may not lead to a clear signal



d Reference set of haplotypes, for example, HapMap

0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

f Testing association at imputed SNPs may boost the signal



a Genotype data with missing data at untyped SNPs (grey question marks)

1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

0	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0
1	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0
...
1	?	?	?	1	?	1	?	0	1	0	?	?	1	?	0
1	?	?	?	1	?	1	?	1	1	1	?	?	1	?	0
...
1	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0
0	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0

e The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

1	1	1	1	1	2	1	0	0	2	2	0	2	2	2	0
0	0	1	0	2	2	2	0	0	2	2	2	2	2	2	0
1	1	1	1	2	2	2	0	0	2	1	1	2	2	2	0
1	1	2	0	2	2	1	0	1	2	2	1	2	2	2	0
2	2	2	2	2	1	2	0	1	2	1	1	2	2	2	0
1	1	1	0	1	2	1	0	1	2	2	1	2	2	2	0
1	1	2	1	2	1	2	0	0	2	1	1	1	2	1	1
2	2	2	1	1	1	1	0	1	2	1	0	1	2	1	1
1	2	2	0	0	2	0	0	2	2	2	1	2	2	2	0

Figure 1.13 Genetic imputation

Genetic imputation is a method of predicting missing genotypes by utilising reference haplotypes

a Genotypes of individuals (rows) for different SNPs (columns), some of which are missing and denoted by a question mark. For a hypothetical biallelic SNP (A/a) there are 3 potential genotypes which are displayed as 0 (aa), 1 (aA) and 2 (AA)

b GWAS association testing does not reveal any significant associations prior to imputation

c Individual genotypes are phased (a statistical process of estimating haplotypes) and this process is displayed for 3 example individuals (the colours denote which reference haplotype they match to in **d**)

d Phased haplotypes are compared to a reference panel of dense haplotypes

e Missing genotypes are imputed using the matched reference haplotypes and this is performed statistically to take account of the uncertainty of each imputed SNP.

f Repeat association testing using imputed SNPs can lead to more significant associations

Figure from (149)

1.5.4 GWAS challenges

There was initial concern that the common variant associations that were discovered by GWAS did not sufficiently explain the heritability that was estimated from pedigree studies, which was termed “missing heritability”.(151) In 2009, GWASs of height had identified ~40 SNP-trait loci which only explained 5% of the variance in height compared to 80% that was estimated from pedigree studies.(151, 152) However, as GWAS sample sizes increased and methodology improved it was recognised that most “missing” heritability in common traits could be explained by a combination of common variants below the genome-wide statistical threshold and to a lesser extent rare variants with larger effect size.(153-155)

A key challenge following a GWAS is to identify causal variants, which are the variants that have a direct or indirect functional effect on disease risk.(124) Whilst linkage disequilibrium enables variation in the genome to be investigated without directly genotyping all SNPs it presents a challenge when associations are discovered. As adjacent variants are correlated and there are more segregating variants than genotyped using SNP micro-arrays, any associated SNP is unlikely to be the causal variant and the association due to LD in the haplotype structure.(124, 137) The process of narrowing down the GWAS association to a causal variant or set of variants is termed fine mapping. Fine mapping is made more challenging as most GWAS associations are in non-coding and inter-genic genomic regions and their functional consequences are more difficult to interrogate than for causal variants within coding sequences in Mendelian diseases.(137, 156) The genetic correlation across common diseases also makes fine mapping more challenging as GWAS associated variants can be associated with multiple traits (pleiotropy).(157) To date, whilst limited studies have bridged the gap between GWAS associations and biological disease mechanisms, progress is now being made as methodologies improve and datasets of functional annotations have become available.(122)

Fine mapping methods for narrowing down causative variants include increasing SNP density, statistical methods (e.g. Bayesian), utilising genomic functional annotations and trans-ethnic fine mapping.(158) Increasing SNP density by either genetic imputation ([Section 1.5.3](#)) or additional genotyping is able to narrow down causal variants by filling in the missing genotypes between SNP micro-array markers.(158)

A number of statistical methods for fine-mapping have been developed including those utilising Bayes theorem, a probability theory based on the probability of prior events.(159) The posterior probabilities generated by a GWAS Bayesian analysis can be used to form “credible sets” of causal variants however, statistical methods alone cannot determine causality.(158, 160) Trans-ethnic fine mapping involves combining separate GWAS for the same trait in different populations and using the difference in LD structure to narrow down causative variants.(158, 161) Most GWAS have a bias towards European populations and a future challenge will be to expand GWAS in other populations, which could uncover novel risk variants and improve trans-ethnic fine mapping.(137)

As most GWAS associations are in non-coding and inter-genic regions of the genome, they may be exerting their effects by influencing transcriptional regulation of genes.(156) In such cases, linking the GWAS association to a biological disease mechanism requires an understanding of the role of non-coding variants in transcriptional regulation. Functions assigned to non-coding variants (annotation) include promoters and enhancers (of transcription), transcription factor binding sites, chromatin accessibility, histone modification and DNase hypersensitivity sites.(158, 162) Genomic annotation has been possible due to the development of genome-wide databases of functional activity including the Encyclopaedia of DNA Elements (ENCODE), Roadmap Epigenomics Project and the Genotype-Tissue Expression (GTEx) project which have mapped regulatory elements across multiple cell types and tissues.(137, 163-165) GWAS associations can be fine-mapped by identifying non-coding variants enriched for regulatory functions from these annotation databases, which can narrow down causative variants for subsequent experimental investigation of functional consequences.(158)

1.5.5 Venous thromboembolism GWASs

Prior to genome-wide associations studies, most VTE genetic risk factors were in genes associated with coagulation and fibrinolysis. Whilst there is strong heritability for VTE, identified genetic risk factors occur in only half of people, suggesting additional genetic associations.(70)

Additional VTE genetic risk factors have been described in several genome-wide association studies (GWAS), a method of identifying common single nucleotide polymorphisms (SNPs) associated with disease ([Section 1.5](#)). A meta-analysis of 7,507 VTE cases from 12 case-control studies and testing ~7 million SNPs, reported 9 significant loci. Of these loci, 6 were involved in the haemostasis / fibrinolytic pathways (*ABO*, *F2*, *F5*, *F11*, *FGG*, *PROCR* genes) and 3 novel loci (*TSPAN15*, *SLC44A2* and *ZFPM2*) were identified ([Figure 1.14](#)).⁽⁷⁰⁾ The *ZFPM2* locus was not replicated in an additional 3009 VTE subjects, but has been identified in a subsequent separate GWAS.⁽¹⁶⁶⁾ The function of the *SLC44A2* and *TSPAN15* loci is unclear, but the locus in *SLC44A2* may be a shared genetic risk factor for coronary artery disease (CAD) and stroke.⁽¹⁶⁷⁾ Other studies have suggested that currently, *ABO* is the only shared genetic risk factor between VTE and CAD.⁽¹⁶⁸⁾ Epidemiological studies have identified older age, smoking and increasing body mass index (BMI) as shared risk factors between VTE and CAD.⁽¹⁶⁹⁾

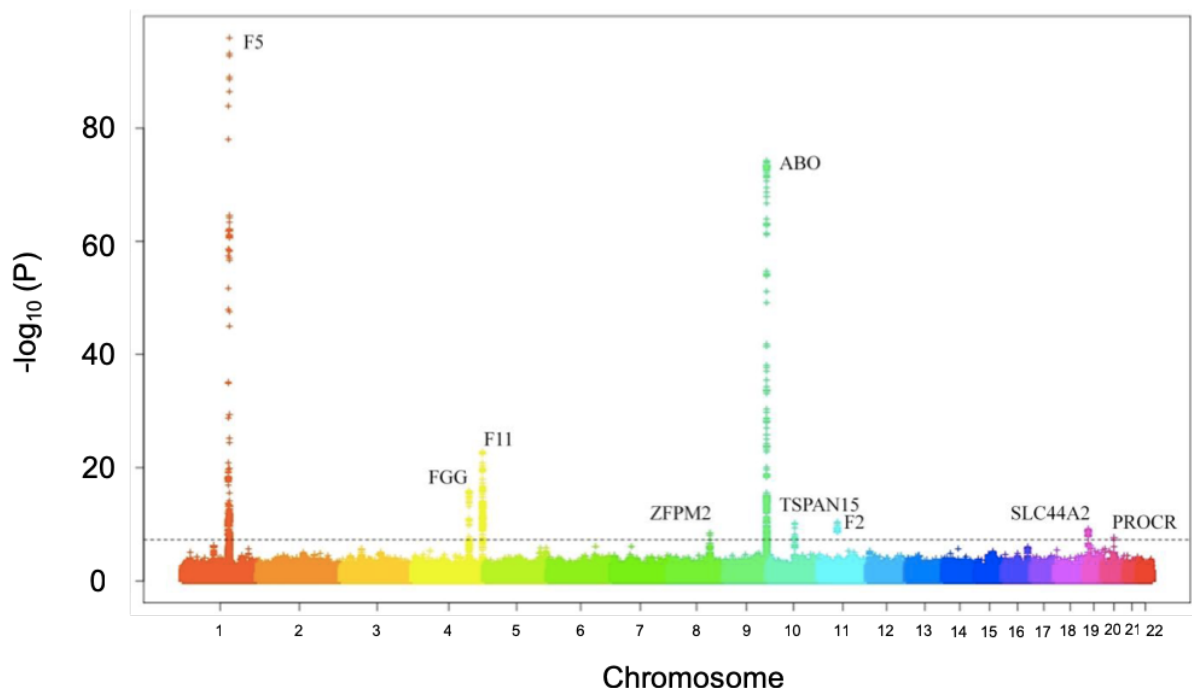


Figure 1.14 Manhattan plot of VTE associated genetic loci

7,507 VTE cases, 52,632 control subjects and 6,751,884 SNPs in the discovery cohort. P -values ($-\log_{10}$) are plotted against genomic position, with those above the dotted line (Bonferroni p -value threshold = 5×10^{-8}) representing significant associations. Figure reproduced © Elsevier, from⁽⁷⁰⁾.

Genetic studies have also been used in several other ways to investigate VTE. An individual's risk of VTE has been investigated by combining multiple genetic risk loci associated with VTE in polygenic risk scores with and without traditional epidemiological risk factors.(170, 171) This method has the potential to be applied to risk stratify patients with PE for the likelihood of developing CTEPH. This is an unmet clinical need where currently no validated tools exist.(172) Genetic studies using Mendelian randomisation, a method to investigate putative causal relationships between modifiable risk factors and disease have found a causal relationship between obesity and VTE.(138, 166, 173)

1.6 Von Willebrand Factor and ADAMTS13

As described in [Section 1.2.2](#), abnormalities in haemostasis are implicated in CTEPH pathobiology.(39, 174) This includes elevated VWF, a multimeric plasma glycoprotein that is synthesized by vascular endothelial cells and megakaryocytes.(35, 175) VWF is stored in Weibel-Palade bodies and alpha-granules, and secreted after activation of the endothelium or platelets.(175) VWF plays an important role in platelet recruitment by mediating adhesion of platelets to the endothelium and is also a carrier protein for the pro-coagulant blood clotting Factor VIII.(175)

VWF activity is normally regulated by ADAMTS13 (a disintegrin and metalloproteinase with a thrombospondin type 1 motif, member 13), a plasma protein that specifically cleaves the more active high molecular weight VWF multimers, thus preventing excessive aggregation of platelets.(176) ADAMTS13 is predominately produced by hepatic stellate cells, in addition to vascular endothelial cells and megakaryocytes.(177) Shear stress causes VWF to undergo a conformational change exposing its active binding site to platelets.(178) This also exposes its A2 domain, which allows ADAMTS13 to bind and cleave ultra-large VWF into smaller and less pro-thrombotic multimers.(178) The mechanism by which ADAMTS13 is regulated has not been fully elucidated. ADAMTS13 is unusual for coagulation enzymes as it is secreted in a constitutively active form and has no known endogenous inhibitor.(177) Plasma proteins that inhibit other members of the ADAMTS family (i.e. alpha-2 macroglobulin) do not affect ADAMTS13 activity towards VWF.(177) It has been proposed that ADAMTS13 is regulated at the VWF substrate level.(177)

1.6.1 Thrombotic thrombocytopenic purpura

The critical role of ADAMTS13 levels in haemostasis is exemplified by thrombotic thrombocytopenic purpura (TTP). TTP is a rare disease that is characterised by microangiopathic haemolytic anaemia, low platelet levels (thrombocytopenia) and microvascular thrombi formation ([Figure 1.15](#)).⁽¹⁷⁹⁾ TTP occurs due to severely reduced plasma levels of ADAMTS13 (activity levels < 10%) that is predominately a result of ADAMTS13 autoantibodies and less frequently due to rare *ADAMTS13* mutations.^(180, 181) The autoantibodies are mainly anti-ADAMTS13 immunoglobulin G (IgG) in three quarters of acute TTP that inhibit the ADAMTS13 mediated proteolysis of VWF.^(179, 180) TTP has a prevalence of ~10 cases per million individuals and the autoantibody form is more common with increasing age, in females and with black ethnicity.^(182, 183) Acute episodes of autoantibody TTP are associated with diseases that increase VWF levels including bacterial infections, other autoimmune diseases (e.g. systemic lupus erythematosus and antiphospholipid syndrome), pregnancy, certain drugs and cancer. However, in 50% no precipitating cause is identified (idiopathic TTP).⁽¹⁷⁹⁾ Acute episodes of TTP can be life threatening and are treated with plasma exchange to remove the ADAMTS13 autoantibodies and/or immunosuppression, with recombinant ADAMTS13 an emerging treatment with ongoing clinical trials.⁽¹⁸⁴⁻¹⁸⁷⁾

Congenital TTP (Upshaw-Schulman syndrome) is caused by rare *ADAMTS13* mutations and occurs in <10% of all TTP.⁽¹⁸⁸⁾ Over 130 heterozygous and homozygous mutations in *ADAMTS13* have been reported and inheritance is primarily autosomal recessive.^(189, 190) Parents of index-cases with congenital TTP do not have clotting abnormalities but do have mildly reduced ADAMTS13 levels consistent with heterozygous *ADAMTS13* variant carriage.⁽¹⁸⁸⁾ Mutations have been described throughout the *ADAMTS13* gene, with 60% being missense mutations and the remaining truncating mutations (nonsense, frame-shift or splice-site mutations).^(189, 190) These mutations cause abnormalities in ADAMTS13 synthesis, activity or secretion.⁽¹⁸⁹⁾

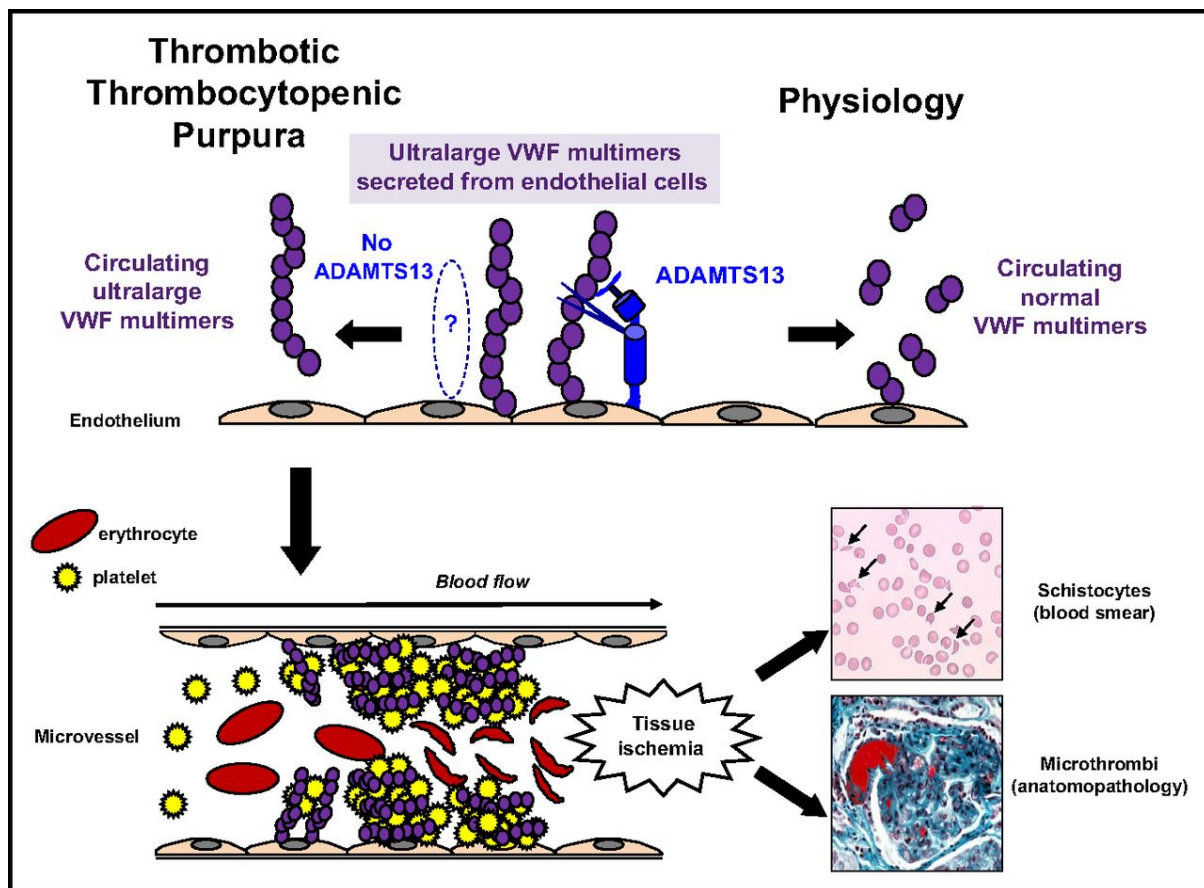


Figure 1.15 Pathophysiology of thrombotic thrombocytopenic purpura

The top right of the figure demonstrates that under normal physiological conditions, ultra-large VWF is secreted from endothelial cells and cleaved by ADAMTS13 into smaller, less pro-thrombotic VWF multimers.

In TTP (top left of the figure) there is a severe reduction of ADAMTS13, leading to much less cleavage of ultra-large VWF multimers and consequently higher circulating levels. The ultra-large VWF bind to platelets that can aggregate within microvessels resulting in tissue ischaemia, platelet consumption (thrombocytopenia) and micro-angiopathic haemolytic anaemia (demonstrated by schistocytes on a blood film).

Figure reproduced © the American Society of Hematology, from (179)

1.6.2 VWF, ADAMTS13 and thrombotic diseases

Plasma VWF is increased in a range of thrombotic conditions including CAD, ischaemic stroke and VTE.(191, 192) Conversely, plasma ADAMTS13 is modestly reduced in CAD and ischaemic stroke.(191, 193) There are discordant findings in

patients with acute PE, with increased, no difference and decreased ADAMTS13 reported.(194-196) VWF and Factor VIII are known to be elevated in CTEPH and do not change following PEA suggesting a role in pathogenesis.(35, 37) Furthermore, the VWF multimer distribution in CTEPH mirrors healthy controls indicating the increase in plasma levels is not just driven by ultra-large VWF.(35) However, the role of ADAMTS13 in CTEPH has not been investigated to date.

VWF and ADAMTS13 have been correlated with d-dimers, a fibrin degradation product that acts as a proxy marker of hypercoagulability.(194, 197) Elevated d-dimer has a well described clinical application in the diagnosis of venous thromboembolism and is also raised in CTEPH. (198, 199)

A large proportion of the variation in VWF levels is genetically determined, with 30% due to *ABO* groups.(200) *ADAMTS13* is situated ~200 kilobases (kb) distal to *ABO* and is genetically regulated with 20% of its variance attributable to common variants at the *ADAMTS13* locus.(201) *ADAMTS13* is not known to vary with *ABO* groups in healthy cohorts.(202) Similar to other thrombotic diseases, the non-O blood groups are over-represented in CTEPH suggesting a mechanism by which VWF levels are increased.(68)

1.7 Pilot CTEPH GWAS data

Royal Papworth Hospital is the UK national referral centre for pulmonary endarterectomy, making it suitably placed to co-ordinate a GWAS, which requires large sample numbers and multi-centre collaboration. In 2014, a pilot CTEPH GWAS was performed by Royal Papworth Hospital and the University of Cambridge.

Provisional GWAS results from 500 patients with CTEPH and 1500 healthy controls identified significant associations in chromosome 9 corresponding to the *ABO* and *ADAMTS13* gene loci. These genes are associated with thrombosis and haemostasis which are plausible pathways involved in CTEPH pathobiology.

1.8 Hypotheses and Aims

The hypothesis of this thesis is that chronic thromboembolic pulmonary hypertension is a polygenic disease with common variant genetic associations. Furthermore, *ADAMTS13* may be a novel genetic association in CTEPH. Functional consequences that have an impact on CTEPH pathobiology may be related to genetic variant associations including dysregulation of the *ADAMTS13*-VWF axis, which will be investigated.

The study aims are:

1. To perform a genome-wide association study in chronic thromboembolic pulmonary hypertension
2. To investigate the *ADAMTS13*-VWF axis in CTEPH patients including its relationship to *ABO* groups and *ADAMTS13* genetic variants
3. To investigate CTEPH sub-phenotype genetic associations

2 Materials and Methods

2.1 GWAS

2.1.1 Sample size calculations

The required GWAS sample size was estimated using the GAS (Genetic Association Study) online power calculator (http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/index.html, accessed 20/1/16).⁽²⁰³⁾ Multiple sample sizes were estimated by varying the SNP effect odds ratio (OR), minor allele frequency (MAF), power, and genetic model (additive, multiplicative, autosomal dominant and autosomal recessive) (**Figure 2.1**). An estimated 1000 cases would be required for a power of 80% to detect an odds ratio of 1.75 assuming a MAF of 0.1 for an additive genetic model.⁽⁹³⁾ In GWAS, the underlying genetic model is unknown and the additive model is most commonly applied to avoid multiple testing.⁽¹⁴⁵⁾ A higher sample size is required to detect associations with lower MAFs and smaller ORs.

2.1.2 Study samples and participants

The study was approved by the regional ethics committee (REC no. 08/H0304/56 and 08/H0802/32) and all study participants provided written informed consent from their respective institutions.

To date, 1555 self-reported Caucasian CTEPH patients have been recruited from 5 European and 1 United States specialist pulmonary hypertension centres. This includes: Bad Nauheim (Kerckhoff Heart and Lung Centre, Bad Nauheim, Germany); Papworth (Royal Papworth Hospital, Cambridge, UK), Imperial (Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK), Leuven (KU Leuven - University of Leuven, Leuven, Belgium), San Diego (University of California, San Diego, USA) and Vienna (Medical University, Vienna, Austria). CTEPH was diagnosed using international criteria.⁽¹⁾ Patients were excluded if they had a PH diagnosis other than group 4 PH. Centres supplied all available bio-banked samples that had been consented for genomic studies and were suitable for DNA extraction. CTEPH samples were compared to 1536

healthy Caucasian controls from the UK Blood Service (UKBS) arm of the Wellcome Trust Case Control Consortium (WTCCC).(204) Shared controls were used in the original WTCCC study and were utilised in the current methodology as the sample numbers were limited by re-genotyping controls ([Section 2.1.3](#)). (204)

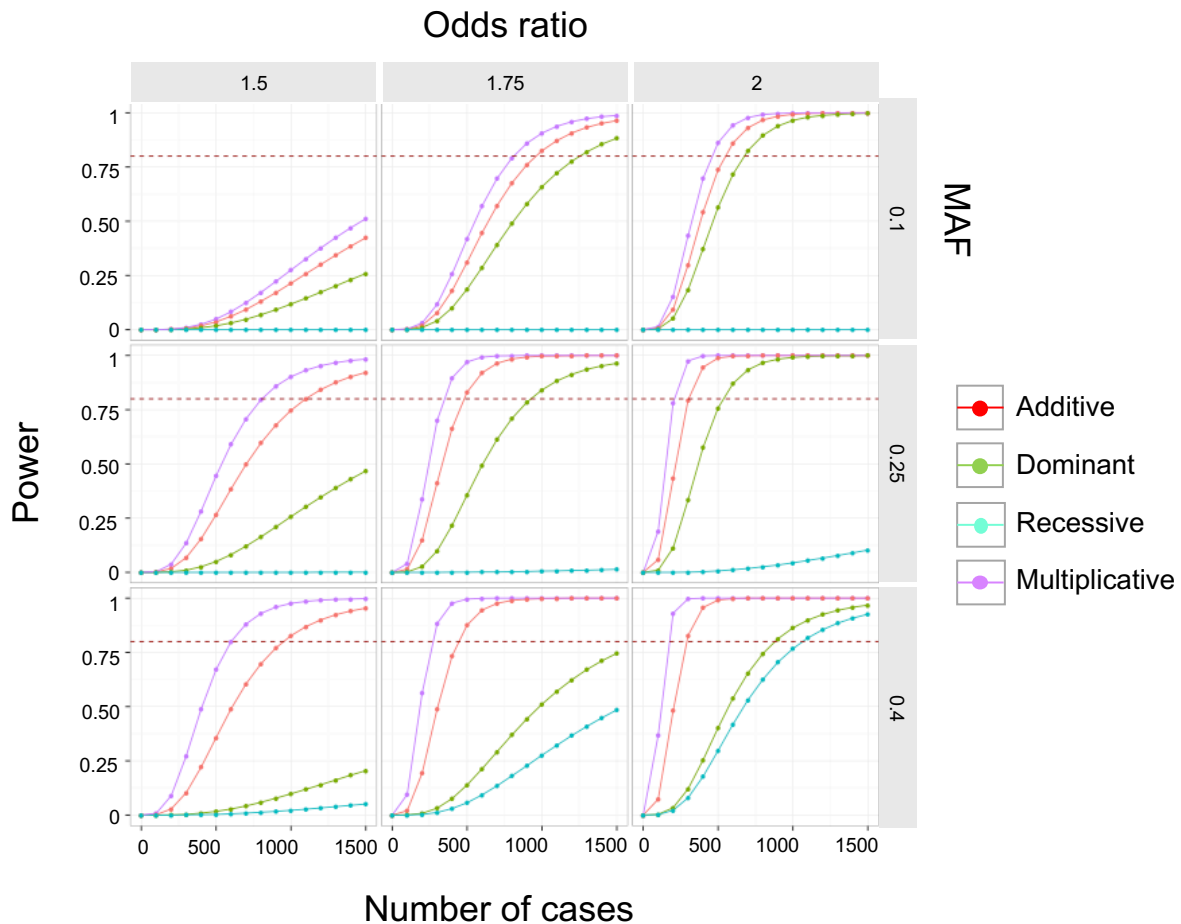


Figure 2.1 GWAS sample size calculations

The estimated number of cases required for a one-stage GWAS when different conditions are varied assuming an equal number of healthy controls: Minor allele odds ratio (top axis: 1.5, 1.75, 2), MAF (minor allele frequency) (right axis: 0.1, 0.25, 0.4), and genetic models (coloured lines). The odds ratio is the ratio of the probability that SNP minor allele is associated with disease to the probability that is not.(205) The genetic models describe the specific relationship between the genotype and phenotype and included additive,

multiplicative, recessive and dominant. For two hypothetical alleles (a and A) at a biallelic SNP locus the three possible genotypes are a/a, a/A and AA. In a multiplicative model the disease risk is increased n-fold for each additional risk allele (e.g. allele A), in an additive model n-fold for a/A and 2n-fold for A/A, in a recessive model two copies of A are required and in a dominant model either one or two copies of allele A for an n-fold increase in disease risk.⁽¹⁴⁵⁾ The dashed red horizontal line represents a power of 80%. Replotted from data obtained from the online GAS power calculator.⁽²⁰³⁾

2.1.3 DNA extraction and DNA microarray

Genomic DNA was extracted and processed from whole blood or buffy coat fractions and quantified with ultraviolet-visible spectrophotometry (LGC, Hoddesdon, Herts, UK). DNA was normalised to a concentration of 50ng/μL and a total volume greater than 4μl (total DNA > 200ng), which was required for the DNA microarray. Genotyping was performed using the Illumina HumanOmniExpressExome-8 v1.2 BeadChip Microarray containing 964,193 single-nucleotide polymorphism (SNP) markers (Kings College, London, UK). The Genome Reference Consortium human genome (build 37) (GRCh37) was used for genomic positions. Three batches were genotyped from 2014-16 (batch1: 2014, batch2: 2015, batch3: 2016). All WTCCC controls were genotyped in batch1 using the same Illumina microarray chip as CTEPH cases.

2.1.4 GWAS quality control

2.1.4.1 GWAS quality control: overview

Sample and SNP quality control (QC) steps are summarised in [Figure 2.2](#). This occurred in three broad stages: micro-array clustering and genotype calling, sample QC and exclusions, SNP marker QC and exclusions. Samples and SNPs were quality controlled in separate micro-array batches and merged together following individual batch QC exclusions. Phasing and imputation were then performed for all remaining samples. Following imputation, additional SNP QC was performed prior to statistical association testing.

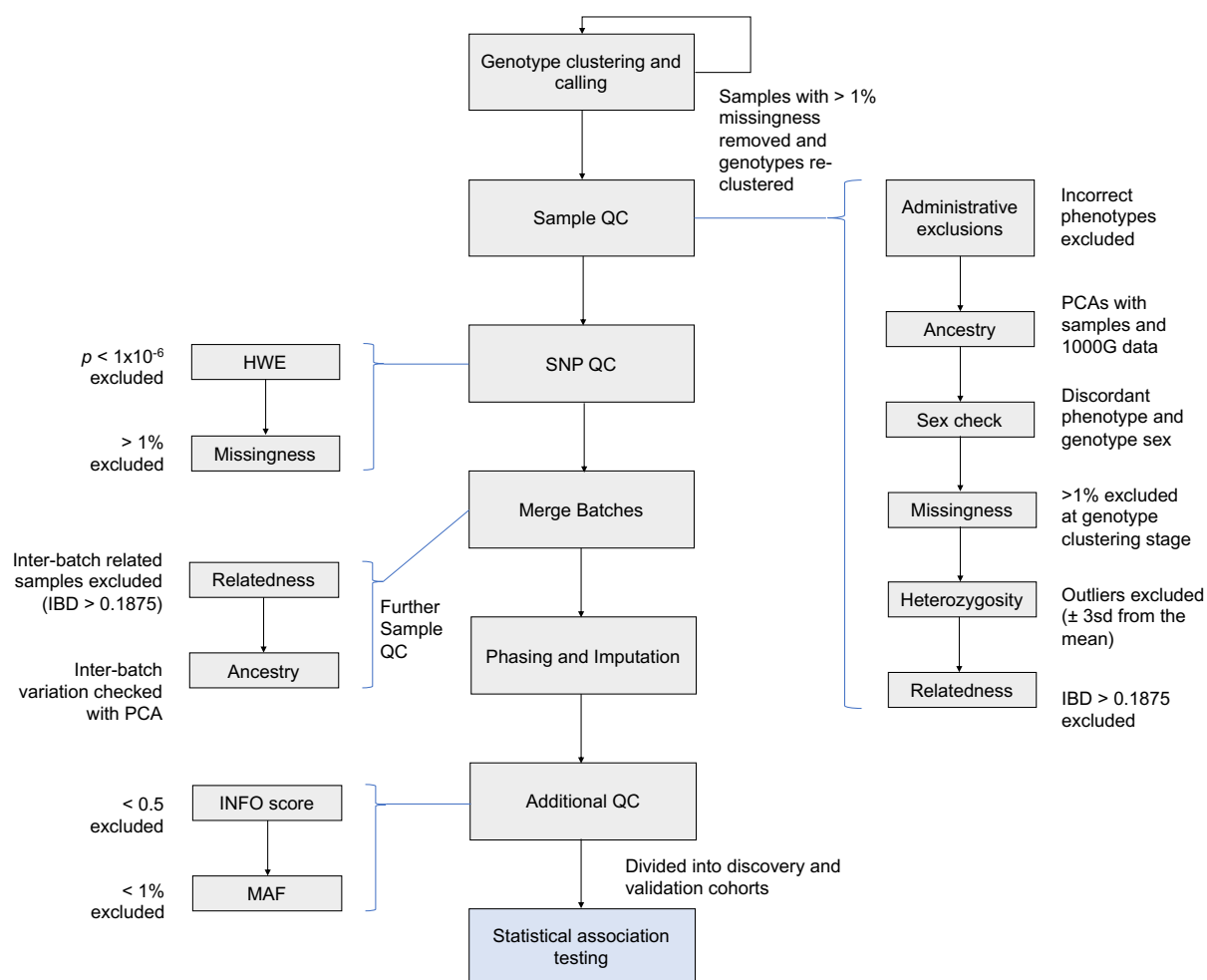


Figure 2.2 Flow chart of GWAS analysis steps

HWE (Hardy-Weinberg equilibrium), IBD (identity by descent), INFO (information score), MAF (minor allele frequency), PCA (principal component analysis), QC (quality control), sd (standard deviation), 1000G (1000 Genomes phase 3)

2.1.4.2 Micro-array intensity data quality control and genotype calling

Each batch of micro-array intensity data was normalised, clustered and genotypes were called independently using Illumina GenomeStudio (v2.0).(206) Samples containing more than 1% missing genotypes (a marker of DNA sample quality) were removed and SNPs were re-clustered to remove confounding from differential SNP exclusions between

batches. The rationale for this approach in relation to the study SNP exclusions is described in [Section 3.2.2.3.2](#). SNPs with poor clustering quality scores (GenTrain score (<0.7) or clustering separation score (<0.5)) were excluded following re-clustering.(206, 207) A higher score (range 0-1; both) reflects better SNP calling quality or separation.

2.1.4.3 Sample quality control

Sample quality control involved an assessment of administrative exclusions (incorrect phenotype), divergent ancestry, relatedness, discordant sex and outlying heterozygosity. Sample genotype missingness exclusions were performed at the genotype calling stage ([Section 2.1.4.2](#)). The sample and SNP QC steps were performed using PLINK (v1.9 beta) unless otherwise stated.(208)

Samples with non-Caucasian ancestry could introduce confounding from genotype differences between (and within) cases and controls that are related to population (ethnicity) differences rather than SNP-trait associations.(147) Divergent ancestry was assessed using principal component analysis (PCA), a statistical dimensionality reduction method. Each batch was merged with the 1000 Genomes phase 3 data using an intersecting set of SNPs.(209) A robust set of independent SNPs ($n=30,609$) was used for PCA using the following criteria: genotype missingness $< 5\%$, SNPs in Hardy-Weinberg equilibrium (HWE) ($p > 1 \times 10^{-5}$), minor allele frequency (MAF) $> 5\%$, independent SNPs (pairwise $R^2 < 0.2$). Furthermore, SNPs in several regions with long-range linkage disequilibrium (LD) were excluded.(210) PCA against all populations in the 1000 Genomes data was conducted using the `SNPRelate` R package, and samples not clustering with European populations were excluded.(211) A second PCA was then performed with the remaining samples against European 1000 Genome reference samples only, and outlying samples were also excluded. Thresholds were set by visual inspection of plots.

Related samples can introduce confounding due to genotype association within families rather than SNP-trait associations. Sample relatedness (including duplicates) was assessed by calculating the proportion of shared alleles at genotyped SNPs for sample

pairs. This was achieved by estimating the parameter identity by descent (IBD) from the calculated parameter identity by state (IBS) in PLINK.(208) To generate the pairwise IBD matrix an independent set of SNPs ($n=228,646$) was generated by LD pruning after removal of long-distance high LD regions.(210) LD was calculated for each SNP pair within a window of 50 SNPs and one of the pair was removed if high LD was present ($R^2>0.2$), before the window was shifted by 5 SNPs and the process repeated. Samples were then excluded if they had a proportion of IBD score ($PI_HAT > 0.1875$), which approximates to a threshold between a 2nd and 3rd degree relative.(212)

Additionally, individual samples were removed due to outlying heterozygosity (3 standard deviations from the mean) and missing genotypes ($>1\%$; described in [Section 2.1.4.2](#)), both markers of DNA sample quality. Sex discordance between centre-reported sex and genotype determined sex was assessed from X-chromosome homozygosity rates but no samples were excluded on this basis (see [Section 3.2.2.2](#) for rationale).

2.1.4.4 SNP quality control

Following sample QC exclusions, SNP markers were removed due to missing genotypes ($>1\%$), deviation from HWE ($p < 1 \times 10^{-6}$), which was only assessed in healthy control samples, differential missingness rate between cases and controls ($p < 1 \times 10^{-5}$) and multi-allelic SNPs. Hardy-Weinberg assumptions are used to estimate allele and genotype frequencies between generations. Deviation from HWE can signify population stratification, genotyping errors or association with the study trait, and hence its application to study controls.(213)

The 3 batches were then merged using an intersecting set of SNPs and further PCA was performed using the robust SNP set previously used for divergent ancestry QC, to check for batch and recruiting centre effects.

2.1.5 Phasing, genetic imputation and post-imputation SNP QC

After quality control exclusions there were 1250 CTEPH cases, 1492 healthy controls and 915,999 SNPs. Phasing and imputation was performed using EAGLE 2 (v2.0.5) and

positional Burrows–Wheeler transform (PBWT) software via the Sanger imputation service (<https://imputation.sanger.ac.uk>, accessed 17/1/17).(214, 215) The reference panel was the Haplotype Reference Consortium (release 1.1), containing ~39 million biallelic SNPs from 32,470 individuals.(216) Following imputation, SNPs were excluded if they had a low minor allele frequency (<1%) or if they were poorly imputed (INFO (information) score < 0.5), with 7,675,738 SNPs remaining for association testing.

2.1.6 Association testing

Case-control association testing was performed using the post-imputation allelic dosages (scale 0-2). Logistic regression assuming an additive genetic model was applied to each SNP marker using PLINK software. Models were adjusted for covariates, namely ancestry informative principal components described in [Section 2.1.4.3](#). The final models used for case-control association testing in the discovery, validation and joint (discovery and validation combined) cohorts used 5 principal components as covariates. Within-case CTEPH analyses also included genotyping batch or recruiting centre as covariates in addition to 5 principal components to investigate any additional confounding.

Genomic inflation was estimated using the parameter lambda, which was calculated by comparing the median of observed and expected test statistics.(145) Genomic inflation values above 1 can indicate population stratification or genotyping errors.(145) To identify independent and secondary signals at associated loci, conditional analysis was performed. Association testing was repeated and conditioned on the allelic dosage of the peak (most significant) SNP in that genomic region. The micro-array intensity clusters for the peak SNP (present pre-imputation) in the associated loci were re-examined to confirm adequate genotyping.

2.1.7 Linkage disequilibrium

Linkage disequilibrium was quantified within the study dataset using PLINK and for a reference dataset (1000 Genomes project data: all European (non-Finnish) populations) using LDlink (<https://analysistools.nci.nih.gov/LDlink/>, accessed 22/01/2018).(217) The

degree of LD was assessed with the parameters R^2 (which takes into account the correlation of SNPs and the allele frequency) and D' (prime).(144)

2.1.8 Genetic ABO groups

The ABO groups A1, A2, B and O were reconstructed using haplotypes from phased data and a described list of tagging ABO SNPs ([Table 2.1](#)). The ABO groups are described in [Section 1.4.2](#) and in addition, the A group can be divided into A1 and A2 subgroups which have very similar chemical structures, with the A1 group expressing more A epitopes (the part of the antigen molecule that the antibody attaches to) and having greater antigenicity.(218) This resulted in 10 groups (A1A1, A1A2, A1B, A1O, A2A2, A2B, A2O, BB, BO, OO), from which blood groups A, B, AB and O were inferred.

The tagging SNPs used to reconstruct the genetic ABO groups A1, A2, B and O from phased haplotypes were: rs8176746, rs8176704, rs687289 and rs507666.(219) The genetic ABO groups were compared to the available ABO antigen groups measured by serology (n=1490 healthy control group) to confirm the accuracy of this method (98% concordance). There were 22 healthy controls and 32 cases that were not able to be classified with an ABO group. The 10 genetic ABO groups were converted into A, B, AB and O groups using the following criteria: A = A1A1, A1A2, A2A2, A1O, A2O; B = BB, BO; AB = A1B, A2B; O = OO.

Genetic ABO group	Haplotype			
	rs8176746	rs8176704	rs687289	rs507666
A1	C	G	A	A
A2	C	A	A	G
B	A	G	A	G
O	C	G	G	G

Table 2.1 Haplotypes used to reconstruct genetic ABO groups

Haplotypes from 4 “tagging” SNPs from phased genotypes were used to assign genetic ABO groups. Table adapted from *Paré et al.* (219)

2.1.9 Fine mapping

Fine mapping was performed to narrow down the associated SNPs and identify a causal variant or a set of variants.(158) Statistical analysis and genomic functional annotations were used in the fine-mapping process.

2.1.9.1 99% credible set

Association testing was performed by Bayesian analysis using SNPtest (v2.5.4-beta3) assuming an additive model with 5 ancestry informative principal components as covariates and using the default priors.(220) The posterior probabilities were then calculated by dividing the Bayes factor for each SNP (within a 200kb region of the peak associated SNP) by the sum of all Bayes factors for that region. Posterior probabilities were ranked in descending order and the SNPs included in the 99% cumulative sum comprised the 99% credible set.(158)

2.1.9.2 Genomic functional annotations

Genomic functional annotations for the associated loci were investigated using the fumaGWAS (Functional mapping and annotation of GWAS) tool (v1.3.2; <http://fuma.ctglab.nl/>).(221) Summary statistics from GWAS association testing were uploaded to the online platform and SNPs were annotated for biological function. SNPs were then mapped to genes based on genomic proximity, cis-expression quantitative trait loci (eQTL) interactions and 3D chromatin interactions.(221) Data from 18 biological repositories was used in these processes including: ANNOVAR (222), Combined Annotation Dependent Depletion (CADD) score (223), RegulomeDB (224), 15-core chromatin state (225), Genotype-Tissue Expression (GTEx)(165), Roadmap Epigenomics Project (164) and Encyclopaedia of DNA Elements (ENCODE) (163). For the fumaGWAS analyses, GTEx (v7) incorporated data from 53 tissue types (including systemic arterial and left heart samples) and ENCODE utilised 127 cell/tissue types, but no resources included pulmonary vascular endothelial or right heart samples.(221)

Independently associated SNPs and correlated SNPs are linked to the GWAS catalog to facilitate interrogation of other trait-SNP associations in the literature.(140)

Association testing was also performed at the gene level using MAGMA (Multi-marker Analysis of GenoMic Annotation) via fumaGWAS.(226) Gene-based analysis may increase power as less statistical tests are performed than when individual SNP markers are tested.(226) SNPs were mapped to 19,311 protein coding genes and then multiple regression (SNP-wise model) was performed with the summary statistics data from the combined GWAS analysis using MAGMA.(226) Gene-set analysis was then performed utilising the gene-based *p*-values for 4728 curated gene sets and 6166 gene ontology (GO) terms from the Molecular Signatures Database (MsigDB v5.2).(221, 227)

2.2 ADAMTS13-VWF axis

2.2.1 Study samples and participants

The study was performed with the same regional ethics committee approval as described for the GWAS (REC no. 08/H0304/56 and 08/H0802/32). All study participants provided written informed consent from their respective institutions.

All consecutive CTEPH patients from the national PEA centre (Royal Papworth Hospital, United Kingdom (UK)) with available plasma samples (August 2013-December 2016) (**Figure 2.3**) and genotype data were included in the study (n=208). CTEPH was diagnosed using international criteria and healthy volunteers (n=68) without major comorbidities were used as a control group (Papworth and Hammersmith Hospital, UK). CTEPH patients were excluded if they had other major contributing factors to their pulmonary hypertension. Additional patient groups were recruited as disease comparators including chronic thromboembolic disease (CTED, n=35), idiopathic pulmonary arterial hypertension (IPAH, n=30) and pulmonary embolism (PE, n=28). CTED was characterised by persistent pulmonary arterial thromboembolic occlusions without pulmonary hypertension (mean pulmonary arterial pressure <25mmHg) in symptomatic patients.(10) The IPAH (Papworth, UK) and PE (Hammersmith, UK) groups were also diagnosed using international criteria.

The 208 CTEPH patients represented 40% (208/514) of all the patients diagnosed with CTEPH during the same period at Royal Papworth Hospital, UK. Healthy controls and disease comparators were selected for the closest possible age- and sex- matching to the CTEPH group, and additionally all IPAH patients were matched for anticoagulation therapy usage but had not had a venous thromboembolism.

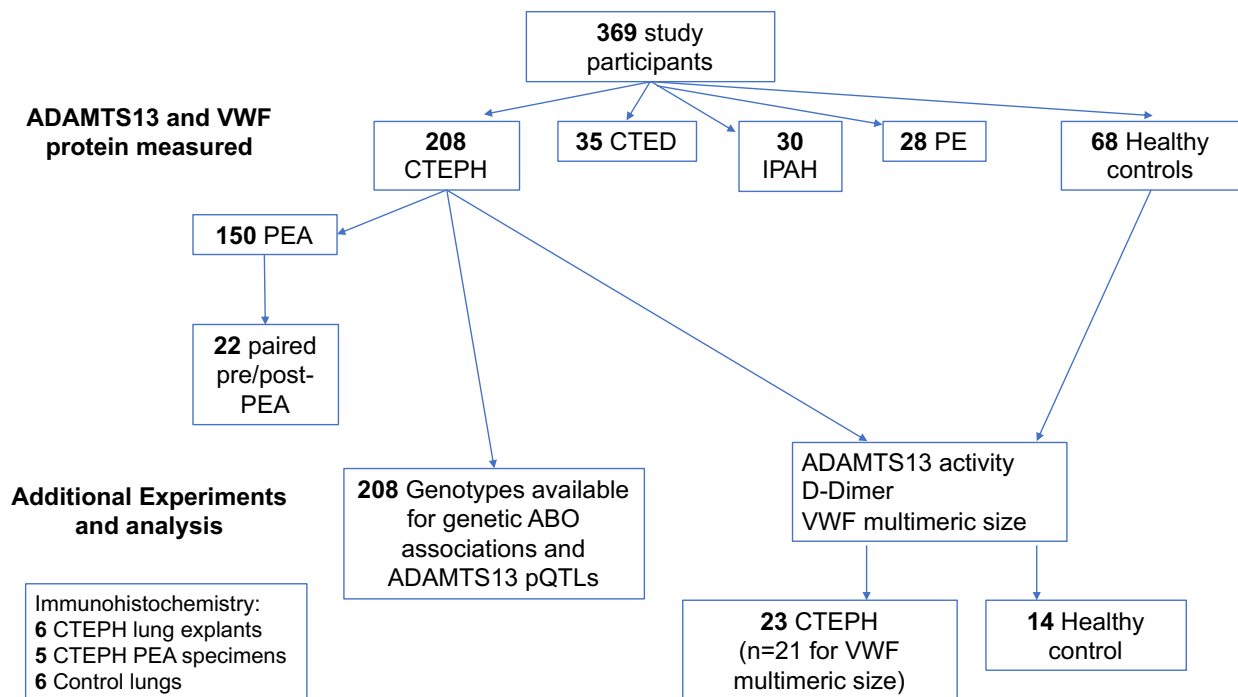


Figure 2.3 Flow chart of study design and study participant numbers

PEA (pulmonary endarterectomy), pQTL (protein quantitative trait loci). The number in each group are shown in bold.

2.2.2 ADAMTS13 and VWF plasma concentrations

Plasma ethylenediaminetetraacetic acid (EDTA) samples were used to measure ADAMTS13 and VWF antigen (Ag) levels by enzyme-linked immunosorbent assays (ELISA). Samples for the CTEPH, CTED and IPAH groups were obtained closest to the time of diagnosis, and pre-operatively for the CTEPH and CTED patients undergoing PEA. Additionally, to assess the effect of PEA, ADAMTS13 and VWF levels were

measured in 22 paired post-PEA samples taken at a follow-up time within 1 year of surgery. The PE group were sampled from a specialist PE follow-up service (Hammersmith, UK) at a median of 220 days following an acute PE.

2.2.2.1 ADAMTS13 plasma concentration

ADAMTS13 plasma antigen levels were quantified using a polyclonal rabbit anti-ADAMTS13 antibody (5µg/mL, anti-TSP2–4 depleted) as previously described.(202, 212) The antibody was immobilised in 96-well microplates (Nunc, Rochester, USA) in 50mM carbonate buffer, pH 9.6 at 4°C overnight. Washes were performed with phosphate buffered saline (PBS) + 0.1% Tween-20 (PBST), and this was repeated between each step. Wells were blocked with 1% bovine serum albumin (BSA) in PBS for 1 hour. Plasma samples were diluted 1:20 using 1% BSA in PBS and added to the wells in duplicate for 2 hours. A standard curve of 0-126 ng/mL was made with normal human control plasma (NHP) (Technoclone, Vienna, Austria) that had known concentrations of ADAMTS13.(202) Bound ADAMTS13 was detected with biotinylated anti-TSP2–4 polyclonal antibody for 2 hours followed by incubation of wells with streptavidin-horseradish peroxidase (HRP) (GE Healthcare, UK) for 1 hour. Plates were developed with a peroxidase substrate (o-phenylenediamine dihydrochloride (OPD); Sigma-Aldrich, Darmstadt, Germany) for 5 minutes and the reaction was stopped with 65µL/well of 2.5M H₂SO₄. Absorbance was read at 492nm (FLUOstar Omega plate reader, BMG Labtech). ADAMTS13 concentrations were obtained by interpolating from the standards fitted with a four-parameter logistic curve. The intra- and inter-assay coefficients of variation were 8 and 12% respectively.

To enable a comparison with other published thrombotic diseases, each ADAMTS13 plasma antigen level was divided by the median of the healthy control group and expressed as a percentage. The CTEPH group was then divided into quartiles of the ADAMTS13 distribution of the healthy control group. The quartile thresholds were used to stratify CTEPH patients and healthy controls into groups using a combination of ADAMTS13 and VWF levels. Odds ratios for the different groups were then assessed using logistic regression adjusted for age, sex, ethnicity and experimental batch.

2.2.2.2 VWF plasma concentration

VWF plasma antigen levels were quantified in a similar well-described manner using a polyclonal rabbit anti-VWF antibody (3.1 µg/mL; Dako, Santa Clara, USA).(212) After overnight antibody immobilisation and washes, wells were blocked with 1% BSA in PBST for 1 hour. Plasma samples were diluted 1:400 in PBST 1% BSA and a standard curve of 0-125 ng/L was made with NHP that had a known concentration of VWF.(202) VWF was detected with 1.1µg/mL Polyclonal Rabbit Anti-Human VWF/HRP (Dako) followed by plate development with OPD for 3 minutes. The intra- and inter-assay coefficients of variation were 5 and 8% respectively.

2.2.2.3 Replicate sample measurements

The ADAMTS13 and VWF ELISAs were performed for all groups in 2016 (batch1). Additional CTEPH samples (n=115) were included in 2017 (batch2) and replicates (ADAMTS13: n=24, VWF: n=12) were used to enable the correction of any batch effects. This was achieved by adjusting batch2 values by the median of the differences if replicates were significantly different between the two batches. If batch adjustment was applied, the validity of this approach was assessed with a multivariable linear model using the uncorrected ADAMTS13 or VWF values.

2.2.3 ADAMTS13 activity, D-dimer and VWF multimeric size

Additional experiments were performed on a subset of the CTEPH (n=23) and healthy control (n=14) groups to identify potential mechanisms for any dysregulation of the ADAMTS13-VWF axis. Plasma lithium heparin samples were used to measure ADAMTS13 activity and D-dimer concentrations. The CTEPH sample subset were those with the lowest ADAMTS13 antigen levels (below the first quartile of the CTEPH group) and the healthy controls were those with ADAMTS13 antigen levels closest to the median of the control group. An estimate of VWF multimeric size was made by measuring VWF collagen binding (VWF:CBA) and comparing this with VWF antigen levels in the CTEPH (n=21) samples with the highest VWF antigen concentrations (above the third quartile of the CTEPH group) and the same healthy control subset.

ADAMTS13 activity was measured with a fluorescence resonance energy transfer (FRET) assay. D-Dimers were quantified by ELISA and VWF multimeric size was assessed using a collagen binding assay.

2.2.3.1 ADAMTS13 activity

ADAMTS13 activity was measured with a fluorescence resonance energy transfer (FRET) assay using a short synthetic VWF peptide (VWF73: PeptaNova, Sandhausen, Germany) containing the ADAMTS13 cleavage site for VWF.(228) Plasma samples and NHP were diluted to 1:10 in reaction buffer (5 mM Bis-Tris, 25 mM CaCl₂ and 0.005% Tween-20 at pH 6.0) in 96-well plates (Nunc, Rochester, USA). FRET-S-VWF73 substrate (an equal volume of 4μM) was added and fluorescence was recorded at 1-minute intervals for 1 hour (FLUOstar Omega plate reader) to monitor substrate proteolysis. Assays were repeated 3 times to obtain the mean fluorescence and ADAMTS13 activity was normalised to NHP, which was defined as 100%.

2.2.3.2 D-Dimer plasma levels

Plasma D-Dimer levels were quantified using an ELISA kit (ab196269, abcam, Cambridge, USA) according to the manufacturer's instructions. Plasma lithium heparin samples from CTEPH patients and healthy controls were used at a dilution of 1:1000.

2.2.3.3 VWF multimeric size

VWF multimeric size was evaluated with a collagen binding assay (CBA) which utilises the increased collagen binding of higher multimeric VWF. Human type III placental collagen (5μg/mL) was immobilised in 96-well microplates plates (Nunc) in 50mM carbonate buffer, pH 9.6 at 4°C overnight. After washes with PBST, wells were blocked with 2% BSA in PBST for 1 hour. Plasma lithium heparin samples were diluted 1:100 in PBST 1% BSA and a standard curve of 0-1000ng/mL was made with NHP. VWF was detected with 1.1μg/mL polyclonal rabbit anti-human VWF/HRP (Dako) followed by plate development with OPD for 3 minutes. Collagen binding is reported as a ratio over the total plasma VWF antigen.

2.2.4 Immunohistochemistry

ADAMTS13 is primarily secreted by the liver and is also produced by vascular endothelial cells however, its expression in pulmonary arteries in health and disease are unclear.(177) ADAMTS13 expression in the pulmonary arteries and PEA specimens was assessed with immunohistochemistry. Tissue sections from peripheral regions of explanted lungs of patients with CTEPH undergoing transplantation (n=6) were compared to tumour-free sections from lung cancer patients undergoing surgical lung resection (n=6). Additionally, ADAMTS13 expression was assessed in the chronic thromboembolic material removed during PEA surgery (n=5). Tissue sections used from explanted CTEPH lungs, controls (tumour-free lung cancer resections sections) and PEAs were a random and representative sample.

Immunohistochemistry is a method for investigating cell or tissue antigens using specific antibodies that can be visualised through staining.(229) The stages of immunohistochemistry include specimen preparation and fixation, antigen retrieval, antibody incubation and washing and counterstaining.(229, 230) Antigen-antibody interactions are identified by immunostaining ([Figure 2.4](#)).

Tissue sections were mounted onto charged adhesive microscope slides (CellPath, UK) and dried overnight at 50°C. Antigen retrieval was performed using a low pH buffer in an automated antigen retrieval system (PT-module, DakoCytomation, UK) following the manufacturers protocol. Primary antibodies raised against polyclonal rabbit anti-ADAMTS13 (1:50 dilution; ab71550, Abcam, USA) and polyclonal rabbit anti-human VWF (1:2000 dilution; ab9378, Abcam, USA) were labelled using dextran-coupled peroxidase (Envision, DakoCytomation, UK), visualised with 3,3'-diaminobenzidine hydrochloride (DAB) to create a brown-coloured reaction product, counterstained with haematoxylin and examined by light microscopy.

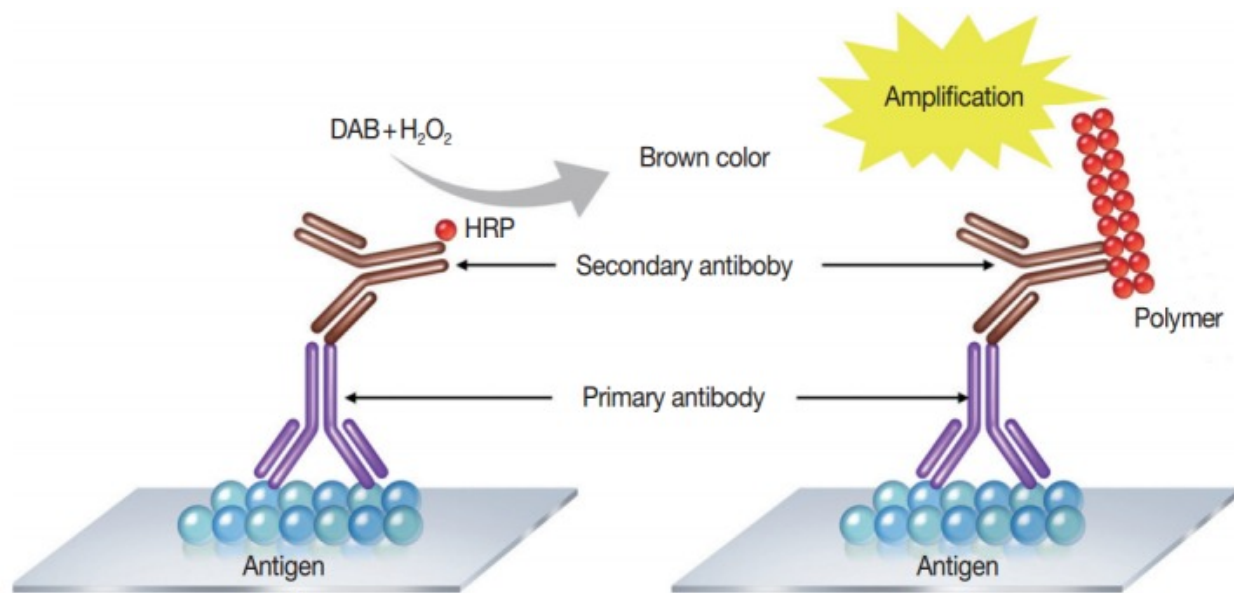


Figure 2.4 Indirect immunohistochemistry using a polymer-based detection system

Immunohistochemistry can be performed using a number of different methods and this is an example of indirect immunohistochemistry that utilises secondary antibodies. The cell (antigen) is exposed to a primary antibody (purple), which binds to a secondary antibody (brown) that is coupled to the enzyme horseradish peroxidase (HRP). This catalyses the DAB (diaminobenzidine) substrate to produce a colour change.

Figure from (230)

2.2.5 Protein quantitative trait loci

208 CTEPH patients with ADAMTS13 / VWF antigen levels and 28 patients with CTED were also included in the CTEPH GWAS. Genotypes were available for 207 (185 CTEPH; 22 CTED) after GWAS quality control exclusions. These patients were included in the analysis comparing ADAMTS13 / VWF protein levels to genetic *ABO* group and the protein quantitative trait loci (pQTL) analysis. Matched genotypes and ADAMTS13 / VWF antigen levels were not available for the healthy control, IPAHA or PE groups.

Associations between the post-imputation allelic dosages of SNPs in the *ADAMTS13* gene \pm 40kb (n=396 SNPs), and log transformed ADAMTS13 protein levels (the

dependent variable) were evaluated using multivariable linear regression. The model was adjusted for age, sex, ADAMTS13 plasma antigen experimental batch and additional models were adjusted for the first 5 ancestry informative principal components used in the GWAS analysis and VWF antigen levels. The *ADAMTS13* \pm 40kb region included the ADAMTS13 cis-pQTLs that have previously been described.(201, 231, 232) A Bonferroni *p*-value threshold $<1.26 \times 10^{-4}$ (0.05/396 variants) was used to denote statistical significance. Partitioning of the variance explained by each variable within the models was performed by averaging over orders using the R package `relaimpo`.(233)

2.2.6 Clinical phenotype data

Phenotype data for the CTEPH, CTED and IPAH groups was recorded closest to the time of diagnosis and pre-operatively for the CTEPH and CTED patients undergoing PEA. This included demographics, haemodynamics, WHO functional class, 6-minute walk distance (6mwd), clinical blood tests, smoking history and anticoagulation therapy usage. Additionally, post-operative haemodynamics were recorded within 1 year of surgery for the CTEPH and CTED patients that underwent PEA, as part of routine care. Haemodynamics were evaluated by right heart catheterisation according to international guidelines and PEA was performed as previously described.(1, 7) The PE group had phenotype data recorded at a follow-up visit (median 220 days) after their acute PE, which also included a ventilation perfusion (VQ) scan to assess residual perfusion defects.

2.2.7 Statistical analysis

The differences in categorical variables between groups were assessed using Chi-squared or Fisher's exact test, and the Cochran-Armitage test for WHO functional class. The differences in continuous variables were assessed using the Mann-Whitney *U* test and the Kruskal-Wallis test. Post-hoc pairwise diagnostic group comparisons were performed using Dunn's test with false discovery rate (FDR) adjustment for multiple testing. For matched values pre- and post-PEA Wilcoxon signed-rank test was used. *P*-values are reported to 3 decimal places and experimental data are reported to 3

significant figures. Data averages are described as median \pm interquartile range unless specified.

Group differences in ADAMTS13 and VWF antigen levels were assessed using multivariable linear regression. ADAMTS13 or VWF plasma levels (dependent variables) were log-transformed after assessing log-likelihoods using the Box-Cox power transformation. Log-transformed ADAMTS13 and VWF were used in all multivariable linear regression models ([Tables 4.4, 4.5, 4.8, 4.9, 4.10, 4.11](#) and [4.12](#)). The models were adjusted for age, sex, experimental batch ([Tables 4.4, 4.5, 4.8, 4.9, 4.10, 4.11](#) and [4.12](#)) and additionally ethnicity ([Tables 4.4, 4.5, 4.9](#) and [4.10](#)), VWF ([Tables 4.4](#) and [4.12](#)) and 5 ancestry informative principal components ([Table 4.12](#)). The β coefficients and confidence intervals are presented as percentage change ($(\exp^{\beta}-1) \times 100$) to enable clinical interpretation of the log-transformed values. Models were checked for normality of residuals, homoscedasticity and multicollinearity (variance inflation factor), with additional checks performed using the R package `gvlma`.(234). Interacting effects between the variables that were used in [Tables 4.4](#) and [4.5](#) were investigated. The significant ($p<0.05$) and informative interactions were included in additional multivariable linear regression models ([Table 4.8](#)).

Spearman's rank correlation coefficients were used to describe associations between ADAMTS13 or VWF protein levels and clinical phenotypes associated with disease severity (pulmonary vascular resistance (PVR), 6mwd and NT-proBNP) and blood markers of inflammation (white cell count (WCC), C-reactive protein (CRP), neutrophil and lymphocyte percentages). *P*-values from correlation testing were corrected for multiple testing using false discovery rate (FDR) adjustment.

2.3 CTEPH phenotype–genotype associations

2.3.1 Phenotype data

A minimal phenotype dataset for CTEPH patients was initially requested from recruiting centres that included age, sex and disease distribution in the pulmonary arteries (surgically accessible = proximal, surgically inaccessible = distal).

Additional CTEPH phenotype–genotype analyses were performed by utilising deeply phenotyped CTEPH datasets from Royal Papworth Hospital and other centres where available. A systematic approach was applied when compiling phenotypes to ensure that data within centres and ultimately across centres was harmonised and reproducible.

2.3.1.1 Data extraction and quality control steps

Exploratory analyses of additional CTEPH phenotype–genotype associations used data almost exclusively from Royal Papworth Hospital. The data extraction and quality control steps described will predominately focus on this centre and is summarised in [Figure 2.5](#).

CTEPH phenotypes were extracted from separate clinical NHS databases held locally at Royal Papworth Hospital. At the time of data extraction (mid-2017), the databases were not linked and searching directly by disease (CTEPH) was not always possible. When a diagnostic search could not be performed, data was extracted using the consultant codes of pulmonary hypertension physicians responsible for CTEPH patient care. The data was then merged with a diagnostic list and filtered to include CTEPH patients ([Figure 2.5](#)).

Each dataset was quality controlled separately and an overview of the size of each starting dataset prior to filtering, and the number of available variables is summarised in [Table 2.2](#).

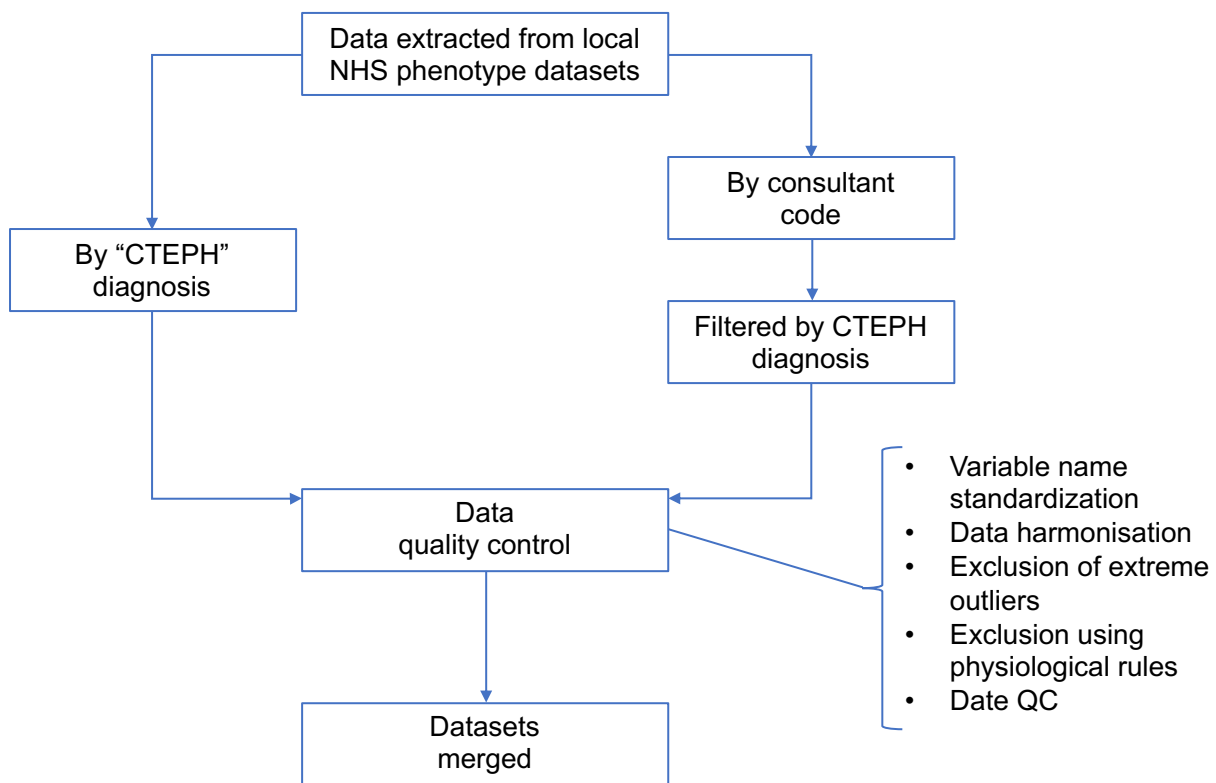


Figure 2.5 Flow chart of phenotype data extraction and quality control

This flow chart applied to the majority of data used for the additional phenotype-genotype analyses from Royal Papworth Hospital. QC (quality control).

Quality control steps were performed in R Markdown annotated documents to facilitate reproducibility.⁽²³⁵⁾ Each dataset had separate QC steps, but generally these involved: variable name standardisation, data harmonisation, exclusion of extreme outliers, exclusion using physiological rules and QC of dates. Data harmonisation involved standardising variable classes (e.g. continuous, characters, factors, dates), standardising data groups within variables, and standardising variable units (of measurement). Data was excluded (set to “NA”) if it was outside a range of biologically/physiologically possible values (e.g. height 5cm) or if inconsistent with physiological rules (i.e. cardiac output is always higher than cardiac index). Dates were standardised and anchor dates (i.e. date of CTEPH diagnosis) were selected to enable the correct assignment of longitudinal data.

	N variables	N total	N individuals
Pulmonary Haemodynamics	30	3,366	2,329
Pulmonary Function tests	20	12,677	3,448
QoL	4	17,896	4,206
WHO FC	2	28,828	4,724
6mwt	7	11,156	4,152
Surgical data	90	1,581	1,578
Clinical blood tests	26	41,744	4,157
Survival	3	6,753	6,750

Table 2.2 Summary of extracted datasets

Summary of a selection of the extracted datasets containing CTEPH phenotypes. Datasets for demographics, co-morbidities, echocardiography and radiology not shown. The number of variables (N variables) for each dataset does not include identification numbers, duplicate variables or multiple dates. The total (N total) number of tests/rows for each dataset is shown together with the number of unique individuals, which differ from N total if there are multiple tests per individual. The surgical dataset primarily contained patients with CTEPH however, the other datasets were extracted by consultant codes or a wider diagnostic group (pulmonary hypertension). For those datasets, the N total and N individuals do not just represent CTEPH patients. Following QC, these datasets were filtered by CTEPH diagnosis. Datasets varied in the time periods covered but generally contained data from 2007-2017. QoL (Quality of life), WHO FC (World Health Organisation functional class), 6mwt (six-minute walking test).

After the QC steps, datasets were merged with a diagnostic list of pulmonary hypertension groups / subtypes and filtered for CTEPH patients. Subsequently, phenotype data was filtered by patients that had been genotyped and included in the GWAS analysis ([Section 3.2.2.2](#)). This enabled additional phenotype-genotype analyses to be performed.

2.3.1.2 Data centralisation

Data centralisation allows disparate datasets from separate tests within a centre ([Table 2.2](#)) and from different centres to be standardised and secured for future use. OpenClinica is an open source clinical data capture and management platform (<https://www.openclinica.com/>) utilised for CTEPH data capture. Electronic case report forms (eCRFs) were designed using OpenClinica templates ([Figure 2.6](#)). The content of the eCRFs was determined by the data availability from extracted CTEPH datasets. Additional parameters were decided by a pulmonary hypertension expert clinical panel that comprised 5 pulmonary hypertension physicians and additional researchers ([Table 2.3](#)). The ultimate aim was to import the extracted CTEPH datasets into OpenClinica and have additional functionality for manual data entry that could be shared between centres.




















Title: Right heart catheter	
Page:	<input type="button" value="Save"/> <input type="button" value="Exit"/> 
Right Heart Catheter	
RHC done?	done * 
Date	<input type="text"/> *  (DD-MMM-YYYY)
Supplemental Oxygen	(select one) * 
Height	<input type="text"/> *  (cm)
Weight	<input type="text"/> *  (kg)
Calculated Values	
Body mass index	<input type="text"/> * 
Body surface area	<input type="text"/> *  (m2)
Heart Rate and Systemic Blood Pressure	
Heart rate measured?	(select one) * 
Systemic BP measured?	(select one) * 
Right atrial and ventricular pressures	
Right atrial pressure measured?	(select one) * 
RVEDP	<input type="text"/> *  (mm Hg)
Pulmonary artery pressure	
PAP (systolic)	<input type="text"/> *  (mm Hg)
PAP (diastolic)	<input type="text"/> *  (mm Hg)
PAP (mean)	<input type="text"/> *  (mm Hg)
Other	
LVEDP	(select one) * 
PAWP (mean)	(select one) * 
Cardiac output	
Cardiac output Method	(select one) * 
Cardiac output	<input type="text"/> *  (l/min)

Figure 2.6 Electronic case report form for right heart catheterisation data

A section of an eCRF for right heart catheterisation (RHC) data is displayed. The eCRF is dynamic with certain sections being hidden/displayed dependent on the question answers. Multiple eCRFs were designed for different phenotype data ([Table 2.3](#))

Phenotype domain	Example eCRF
ID	Centre ID, Research ID, Genotype ID
Demographics	Sex and ethnicity
Symptoms and risk factors	CTEPH symptoms, co-morbidities, VTE risk factors, VTE management, CTEPH risk factors
Diagnosis	MDT diagnosis
Investigations	6mwt, clinical blood tests, RHC, PFTs,
Management	Medications, PEA, BPA
Outcome	Mortality, longitudinal outcome data (RHC, WHO FC, QoL, 6mwt)

Table 2.3 Phenotype domains and example eCRFs for OpenClinica data capture

The CTEPH phenotype groups and example eCRFs for each domain.

ID (identification), VTE (venous thromboembolism), MDT (multi-disciplinary team), 6mwt (six-minute walking test) RHC (right heart catheterisation), PFTs (pulmonary function tests), PEA (pulmonary endarterectomy), BPA (balloon pulmonary angioplasty), WHO FC (World Health Organisation functional class), QoL (quality of life).

2.3.2 GWAS associations

2.3.2.1 Additional case-control analysis

Case-control analysis was performed as described in [Section 2.1.6](#). The loci associated with VTE were examined in the CTEPH case-control GWAS to identify differential associations.(70)

Abnormalities in haemostasis and fibrinolysis are implicated in the pathobiology of CTEPH. Patients with VTE and CTEPH are treated with anticoagulation, which is predominantly the drug warfarin. GWASs have identified loci associated with warfarin metabolism.(236) In the CTEPH GWAS, these loci were examined to establish whether inadequate anticoagulation due to genetic variants related to warfarin metabolism were associated.

Post-imputation allelic dosages were converted to alleles (A, T, C, G) using PLINK. Genotypes for the significant case-control GWAS associations ($p < 5 \times 10^{-8}$) were used in a logistic regression model with case/control as the dependent variable. Genotypes for the lead SNP associations were also used in an analysis of CTEPH disease severity. Disease severity at baseline (closest to diagnosis) was assessed by right heart catheter haemodynamics (mPAP, CI, PVR) for the CTEPH group. Haemodynamics were also stratified by genetic *ABO* groups ([Section 2.1.8](#)) and additional CTEPH severity measures (6mwd and WHO FC) were stratified by *ABO* group.

Survival was investigated in CTEPH patients from Royal Papworth Hospital following PEA as this represented the largest group that received the same intervention. Survival from the time of PEA was recorded until April 2018 using a centralised national resource. Post-PEA survival differences between genetically inferred *ABO* groups were assessed using Kaplan-Meier plots. As multiple variables can influence post-PEA survival in CTEPH, a cox proportional hazards model was constructed with age, sex and pre-operative disease severity (mPAP) as covariables. The cox models were checked for proportional hazards assumptions, influential observations and non-linearity.

2.3.2.2 Additional phenotype-genotype associations

CTEPH can occur in different pulmonary artery distributions from the central, proximal pulmonary arteries to the distal vasculature. Different risk factors have been associated with distal and proximal CTEPH and this may reflect differing pathobiological mechanisms.(11, 78) Disease distribution (proximal (operable) or distal (surgically inaccessible) disease) was used as the dependent variable in a logistic regression analysis within CTEPH cases. SNP allelic dosages, 5 ancestry informative principal components ([Section 2.1.4.3](#)), age, sex and recruiting centre were included in the model as independent variables.

Distal CTEPH and post-PEA persistent or recurrent PH have been grouped together for clinical drug trials.(24) The rationale is that post-PEA PH may be due to residual distal disease, as the more proximal chronic thromboembolic material has been removed. A

separate logistic regression analysis was performed comparing proximal CTEPH with distal CTEPH and post-PEA residual PH.

Patients can have a comparable amount of chronic thromboembolic obstruction but differing degrees of pulmonary hypertension and right ventricular adaptation.(237) Pulmonary and right ventricular adaptation may be associated with different genetic associations within CTEPH. Linear regression was performed with either mPAP, CI or PVR as the dependent variable and the independent variables included SNP allelic dosages, 5 ancestry informative principal components, age, sex and recruiting centre.

2.4 Software and online tools

The analyses were performed using the following software and online tools: GAS online power calculator(203), Genomestudio (version(v)2.0)(206), PLINK (v1.90beta)(208), bcftools (v1.4.1)(238), LDLink(217), Snptest (v2.5.4-beta3)(220), fumaGWAS (v1.3.2)(221), LocusZoom (v0.4.8)(239), gnomAD(240), Ensembl 37 (241), R (v3.4.3)(235) and RStudio (v1.1.414)(242). The R packages used included: MASS(243), coin(244), gvlma(234), PMCMR(245), SNPRelate(211), relaimpo(233), jtools(246), forestmodel(247), forestplot(248), survival(249), survminer(250) and the tidyverse suite(251).

3 GWAS

3.1 Introduction

Investigating CTEPH aetiology and pathobiology has been challenging for several reasons. Firstly, no small animal model exists that adequately recapitulates failure of thrombus resolution or the chronic pulmonary artery and right ventricular changes observed in CTEPH. Secondly, investigating biological pathways related to haemostasis and fibrinolysis is confounded by anticoagulation treatment that all CTEPH patients receive. Finally, CTEPH is an uncommon complication of PE affecting ~3% of PE survivors. Therefore, very large cohorts of extensively phenotyped individuals post-PE would be required to establish the genetic and environmental drivers of CTEPH development.

GWAS has the potential to circumvent some of the challenges to investigating the aetiology of CTEPH ([Section 1.2.2.7](#)). A case-control GWAS was performed to compare SNP allele frequencies between CTEPH patients and healthy controls. The statistically significant associations will guide subsequent investigation of CTEPH pathobiology. As three-quarters of CTEPH patients have had a preceding pulmonary embolism the ultimate aim is to identify unique and differential genetic associations from resolved PE. This may also inform and enhance clinical risk prediction post-PE.

Royal Papworth Hospital and the University of Cambridge are optimally positioned to perform and co-ordinate this multi-centre international study. Royal Papworth Hospital is the National referral centre for pulmonary endarterectomy, the surgical operation to remove chronic scarred blood clots from the proximal pulmonary arteries of CTEPH patients. An extensive biobank of CTEPH samples has been established that have been utilised in the CTEPH GWAS.

The aim of this Chapter was to perform a GWAS in CTEPH to identify genetic associations with disease susceptibility.

3.2 Results

3.2.1 Study samples and participants

1555 CTEPH cases and 1536 healthy control samples were recruited and genotyped from 6 international centres (**Figure 3.1A**). CTEPH samples were genotyped in 3 batches from 2014-2016 (**Figure 3.1B**). All healthy controls samples (WTCCC) were genotyped in batch1 (**Figure 3.1C**). Royal Papworth Hospital was the largest recruiting centre (n=747) and together with Hammersmith Hospital, Imperial Healthcare NHS trust, made up the discovery cohort (n=841). The remaining European and US centres comprised the validation cohort (n=714).

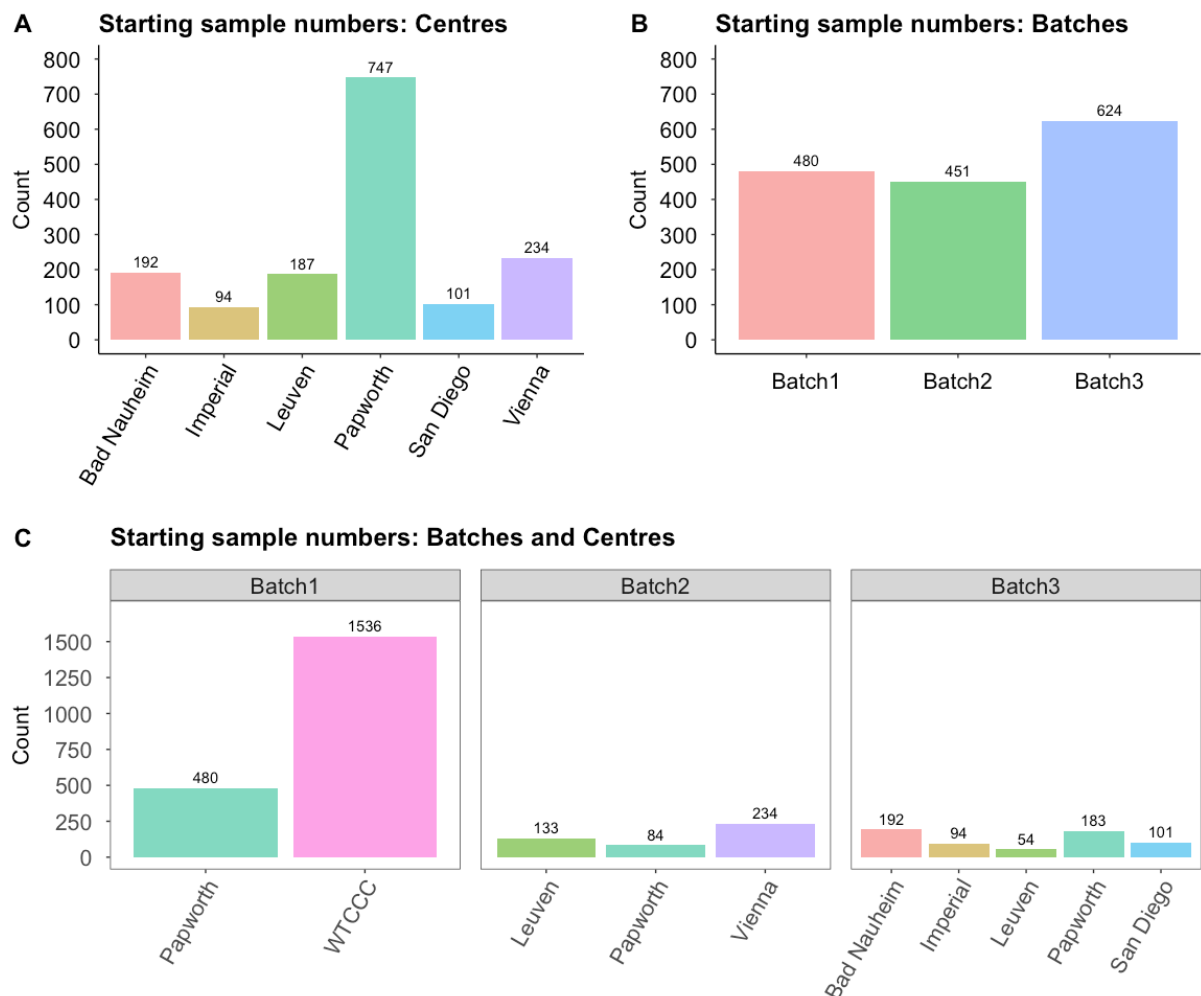


Figure 3.1 GWAS sample numbers prior to exclusions

A Starting sample numbers for CTEPH cases by centre and **B** genotyping batch (batch1: 2014, batch2: 2015, batch3: 2016).

C Starting sample numbers for CTEPH cases and healthy controls (WTCCC) by centre and batch.

n displayed at the top of each count bar. Bad Nauheim (Kerckhoff Heart and Lung Centre, Bad Nauheim, Germany); Papworth (Royal Papworth Hospital, Cambridge, UK), Imperial (Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK), Leuven (KU Leuven - University of Leuven, Leuven, Belgium), San Diego (University of California, San Diego, USA), Vienna (Medical University, Vienna, Austria), WTCCC (Wellcome Trust Case Control Consortium, UK).

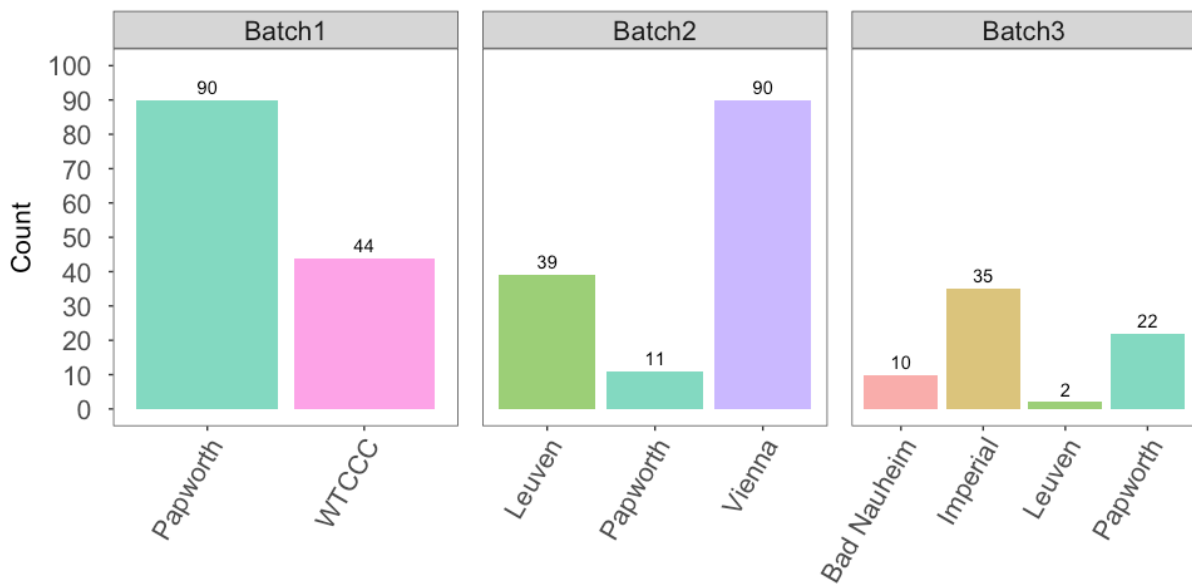
3.2.2 Study exclusions and GWAS quality control

3.2.2.1 Sample exclusions: overview

A total of 349 samples (CTEPH cases = 305 (20%), healthy controls = 44 (3%)) were excluded from the study ([Figure 3.2](#)). Samples were excluded on the basis of thresholds and criteria defined in the methods ([Section 2.1.4.3](#)) for ancestry, heterozygosity, relatedness, sample genotype missingness and administrative (i.e. incorrect disease phenotype) ([Figure 3.2](#)). The relatively high CTEPH case exclusions of 28% (n = 134) for batch1 and 31% (n=140) for batch2, improved to 11% (n = 69) for batch3 ([Figure 3.2B](#)).

The number of exclusions from the healthy control WTCCC group (n=44) was similar to other studies which have used the same healthy control cohort.(204)

A Sample exclusions: Batches and Centres



B Sample exclusions: Batch

Exclusion reason	Batch1	Batch2	Batch3
Ancestry	51	19	60
Heterozygosity	30	10	12
Relatedness	54	2	7
Missingness	20	41	5
Administrative	6	81	2
Total (unique)	134	140	69

C Sample exclusions: Centre

Exclusion reason	Bad Nauheim	Imperial	Leuven	Papworth	Vienna	WTCCC
Ancestry	7	34	4	65	6	14
Heterozygosity	3	8	2	23	5	11
Relatedness	2	2	0	34	2	23
Missingness	2	0	36	18	7	3
Administrative	0	0	0	12	77	0
Total (unique)	10	35	41	123	90	44

Figure 3.2 GWAS sample exclusions

A Sample exclusion numbers for CTEPH cases and healthy controls (WTCCC) by centre and batch

B Sample exclusions for CTEPH cases by genotyping batch and **C** centre.

The "Total (unique)" row in **B** and **C** represents the distinct individual sample exclusions. As samples could be excluded on more than one criterion, the individual exclusions do not equal the "Total (unique)" values. Administrative reasons were due to either the incorrect disease phenotype (non-CTEPH) or non-Caucasian (self-reported) samples that were identified after micro-arraying. There were no exclusions for the San Diego centre (not shown).

n displayed at the top of each count bar. n=343 exclusions (unique) in **Figure 3.2A-**

C There was an additional n=6 exclusions when batch1, batch2 and batch3 were merged due to relatedness, resulting in n=349 total GWAS exclusions.

3.2.2.2 Sample exclusions: GWAS quality control

Samples for exclusion from the study were identified by quality control steps outlined in [Figure 2.2](#) (Materials and Methods). Following genotype calling, samples with a genotype missingness of >1% were excluded (n: batch1=20, batch2=41, batch3=5). The rationale for this approach is described in [Section 3.2.2.3.2](#). Subsequently, each sample quality control step was performed on all the remaining samples prior to further sample removal.

Firstly, samples were identified and excluded due to administrative reasons (incorrect phenotype or self-reported ethnicity) (n: batch1=6, batch2=81, batch3=2).

Principal component analysis using a robust set of independent SNPs was used to identify samples with outlying ancestry (n: batch1=51, batch2=19, batch3=60) ([Figure 3.3](#)). Samples were initially excluded if they did not cluster with super-populations from 1000 genomes data ([Figure 3.3A, D and G](#)). PCA was then repeated, and samples that did not cluster with 1000 genomes European populations were excluded ([Figure 3.3B, E and H](#)). The remaining CTEPH cases and healthy controls cluster with the 1000 genomes European populations ([Figure 3.3C, F and I](#)).

Sample relatedness (including duplicates) was assessed by estimating identity by descent in PLINK. Samples were excluded if they had a proportion of IBD score (PI_HAT) > 0.1875, which approximates to a threshold between a 2nd and 3rd degree relative (n: batch1=54, batch2=2, batch3=7) ([Figure 3.4](#)). In the WTCCC healthy control group there were 23 individual samples with IBD scores above the QC threshold (only half of which would have been excluded), which is similar to other studies that have used the same healthy control cohort.(204)

DNA quality was assessed by excluding samples with outlying genotype missingness (>1%) and outlying heterozygosity ([Figure 3.5](#)). The genotype missingness exclusion step was previously described and is performed after genotype calling. Samples with high or low heterozygosity rates can represent contamination or inbreeding respectively and were excluded if they were ± 3 standard deviations from the mean of the batch group (n: batch1=30, batch2=10, batch3=12).(142)

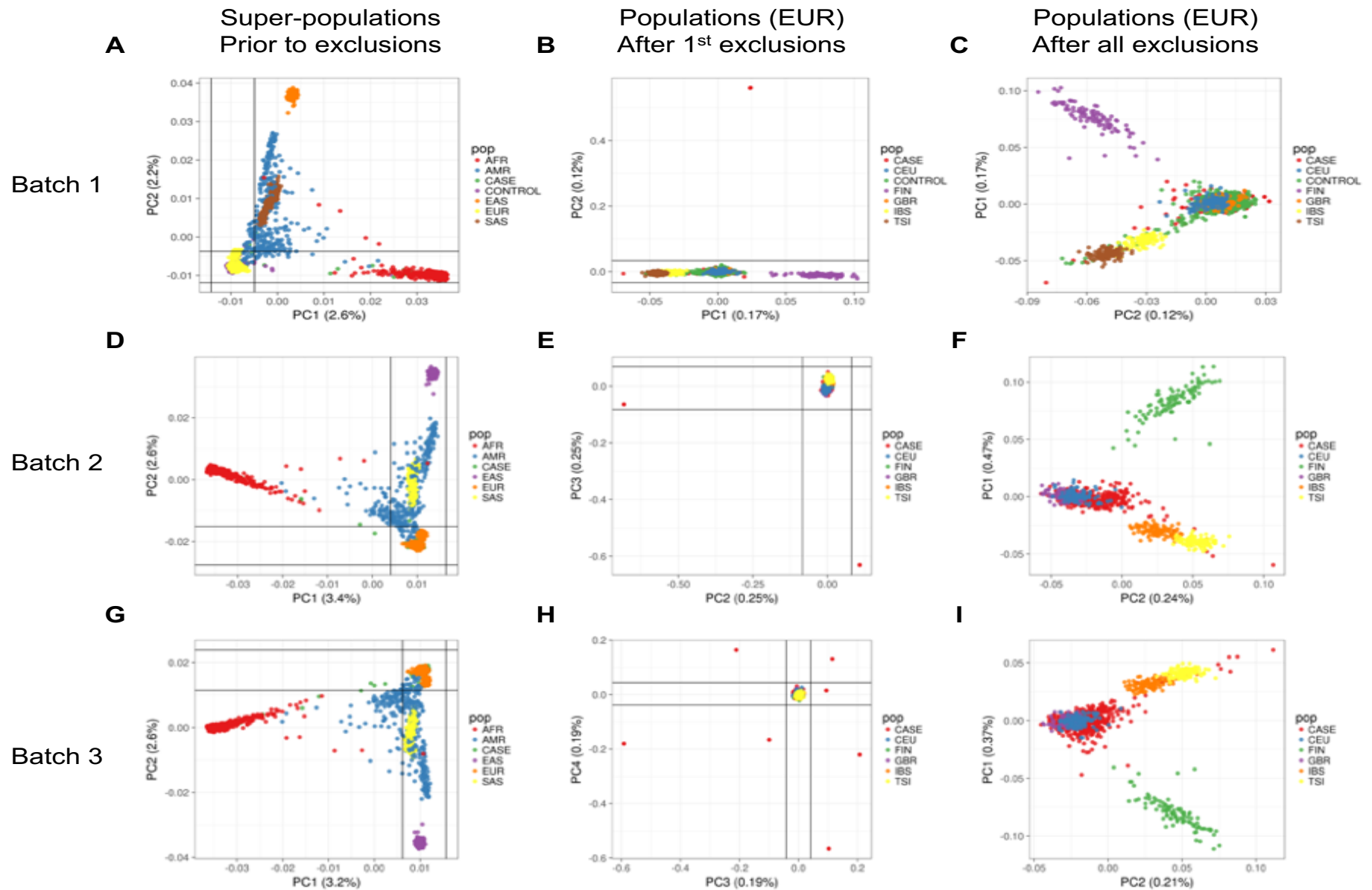


Figure 3.3 Divergent ancestry assessed by principal component analysis

Batch1 (A-C) contains both CTEPH cases and healthy control samples, and Batch2 (D-F) and Batch3 (G-I) contain CTEPH cases only.

The first column of plots (A, D and G) are PCAs including the 1000 genomes super-populations (AFR, AMR, EAS, EUR and SAS). The second column of plots (B, E and H) are PCAs including the 1000 genomes European (EUR) populations (CEU, FIN, GBR, IBS and TSI). The third column of plots (C, F and I) are the PCAs after samples with divergent ancestry have been excluded. The horizontal and vertical black lines in A, B, D, E, G and H are the clustering exclusion thresholds, set by visual inspection. Group colours vary across plots and are defined in the individual legends.

Study samples: CASE (CTEPH cases), CON (WTCCC healthy controls)

Super populations: AFR (African), AMR (Admixed American), EAS (East Asian), EUR (European), SAS (South Asian).

Populations: CEU (Utah Residents (CEPH) with Northern and Western European Ancestry), FIN (Finnish in Finland), GBR (British in England and Scotland), IBS (Iberian Population in Spain), TSI (Toscani in Italia).

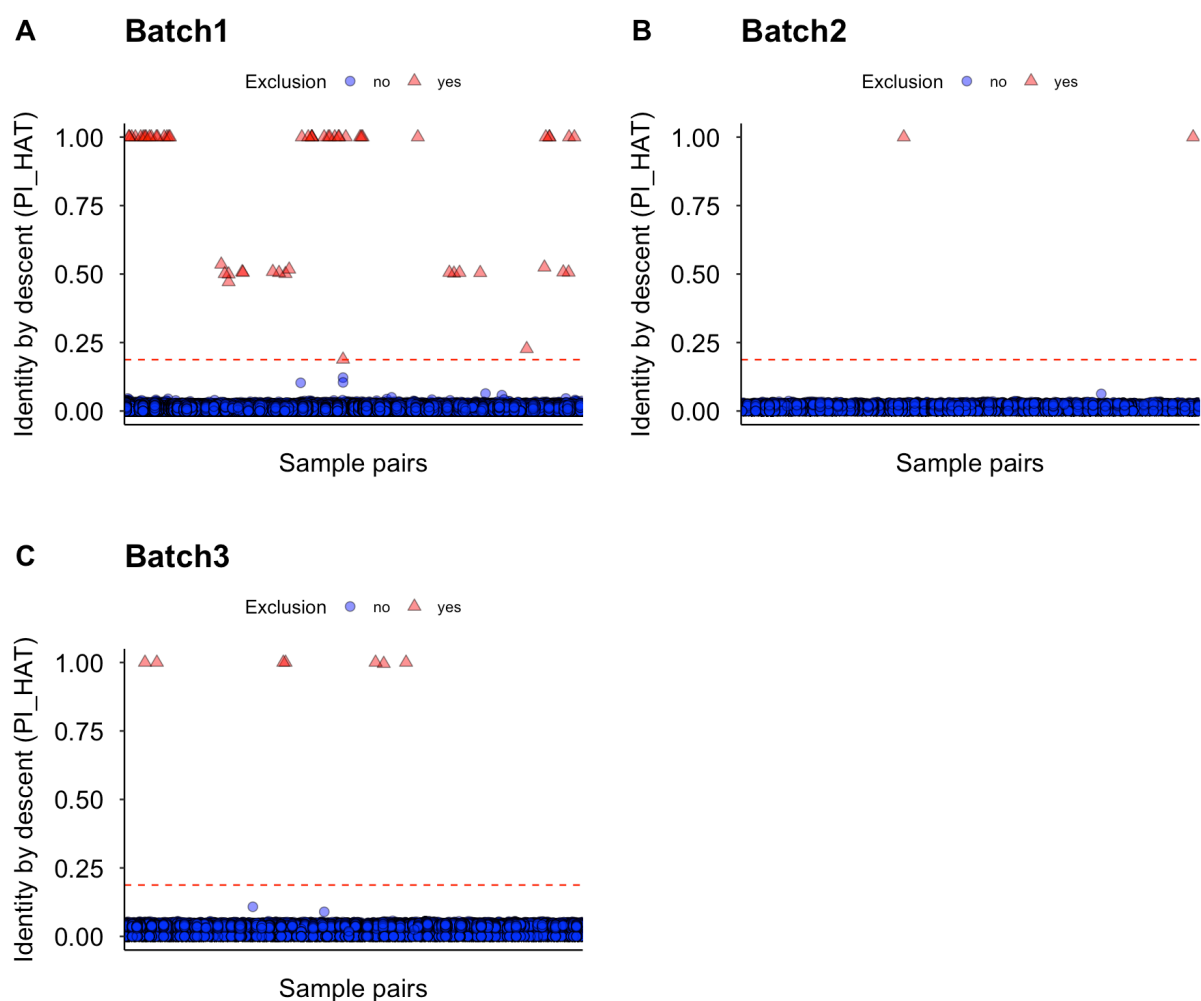


Figure 3.4 Sample relatedness assessed by identity by descent

The proportion of identity by descent (IBD) between pair-wise samples was determined in PLINK using the PI_HAT parameter. PI_HAT ranges from 0 (unrelated) to 1 (duplicate or monozygotic twins). PI_HAT is calculated by: $P(\text{IBD}=2) + 0.5 \times P(\text{IBD}=1)$, where $P(\text{IBD}=2)$ and $P(\text{IBD}=1)$ are the probabilities that at a given locus 2 or 1 alleles respectively are identical by descent.(207) Sample pairs are plotted on the x-axis (Batch1=1,983,036, Batch2=55,955, Batch3=190,036). The dotted red horizontal line represents a PI_HAT threshold of 0.1875 with excluded samples shown by red triangles. For a pair of samples with high PI_HAT values, the sample with the lowest sample genotype rate was excluded.

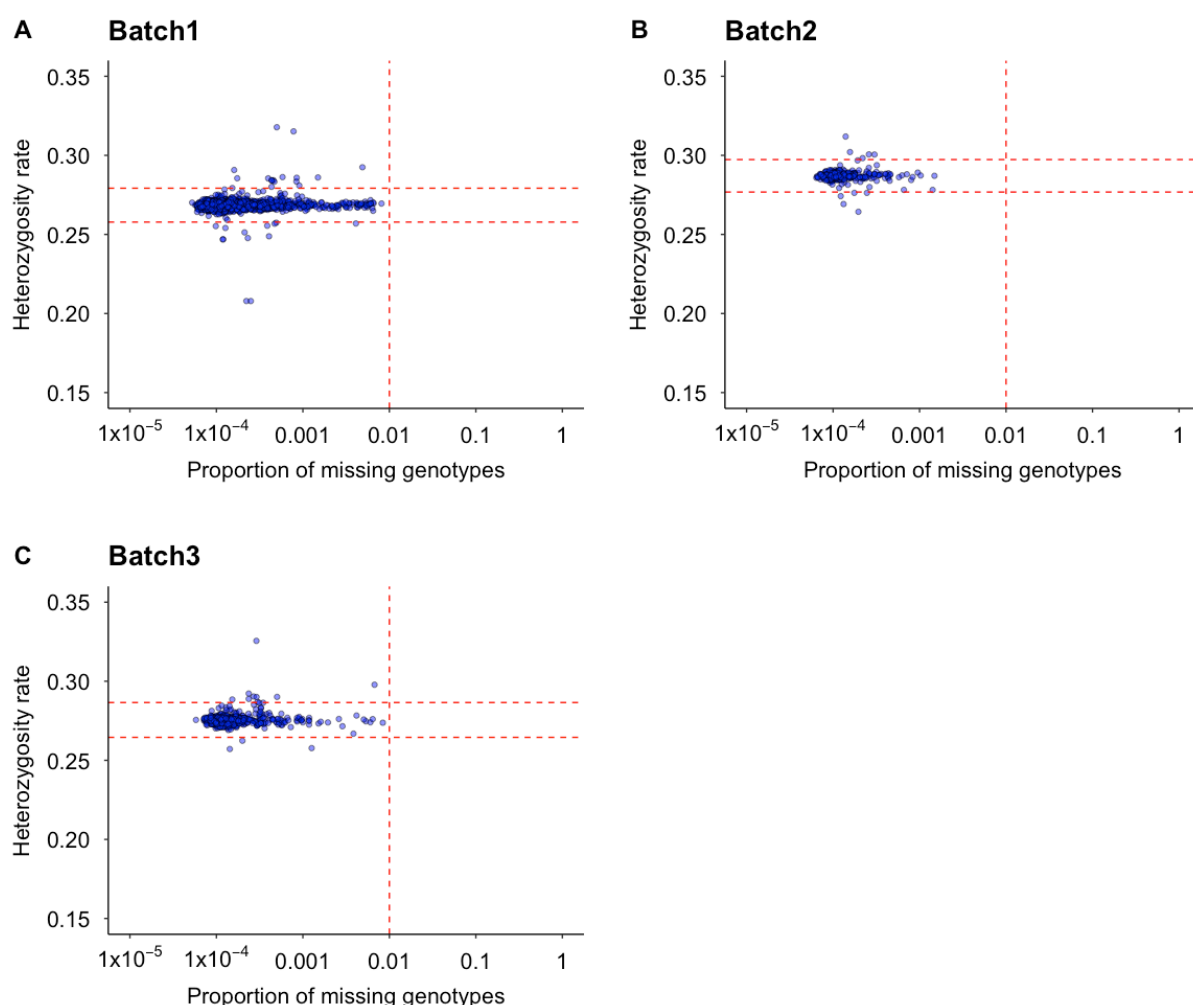


Figure 3.5 Outlying sample heterozygosity plotted against sample genotype missingness

The autosomal heterozygosity rate and genotype missingness were calculated using PLINK. The dotted red horizontal and vertical lines represent thresholds of a heterozygosity rate ± 3 standard deviations from the mean and a genotype missingness of 1% respectively. No sample outliers for genotype missingness are displayed as they were removed during an earlier QC step. Missingness values were converted to \log_{10} to improve visualisation.

Sex discordance between centre-reported sex and genotype determined sex was assessed from X-chromosome homozygosity rates using PLINK ([Figure 3.6](#)). PLINK assigns a male or female genotype if the X-chromosome homozygosity estimate is >0.8 and <0.2 respectively.(208) There were 19, 1 and 8 samples that had discordant

sex in batches 1,2 and 3. Sex discordance may arise for legitimate reasons in addition to sample handling and identification problems. Patients with a self-reported gender different to that assigned at birth may be receiving oestrogen therapy and at increased risk of venous thromboembolism and potentially CTEPH. To retain sample numbers, disease phenotypes were confirmed separately with centres and no exclusions were made due to discordant sex.

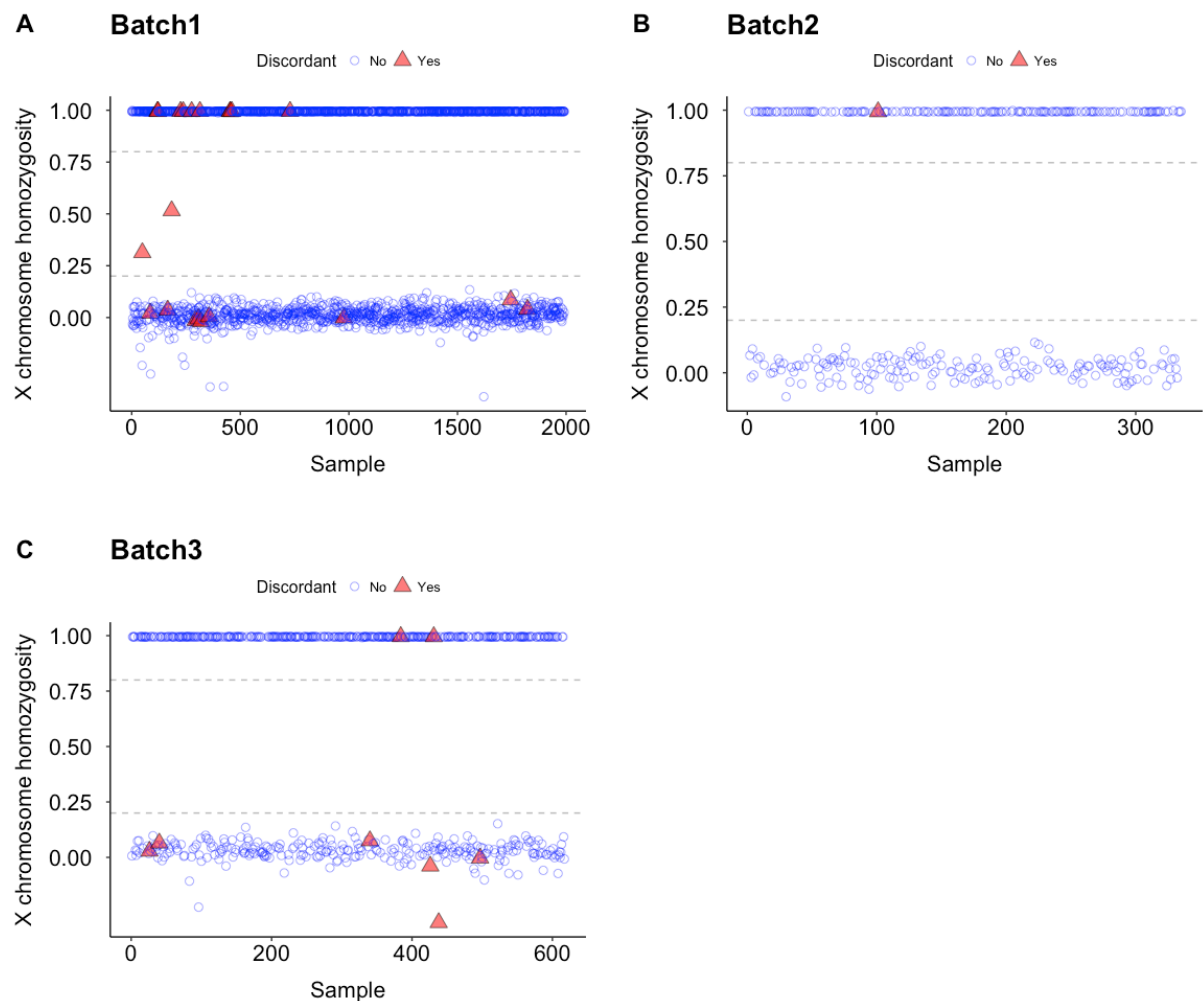


Figure 3.6 Discordant sex for individual samples

Discordant sex was assessed using X-chromosome homozygosity calculated in PLINK. The dotted horizontal grey lines represent thresholds of 0.8 and 0.2 for male and female genotype assignment respectively. The red triangles represent individual samples with discordant sex. An X-chromosome homozygosity score of between 0.2-0.8 was called as missing genotype sex.

Following quality control exclusions, 1250 CTEPH cases and 1492 healthy controls remained for genetic imputation and statistical association testing. The number of individual samples for each centre and genotyping batch are summarised in [Figure 3.7](#).

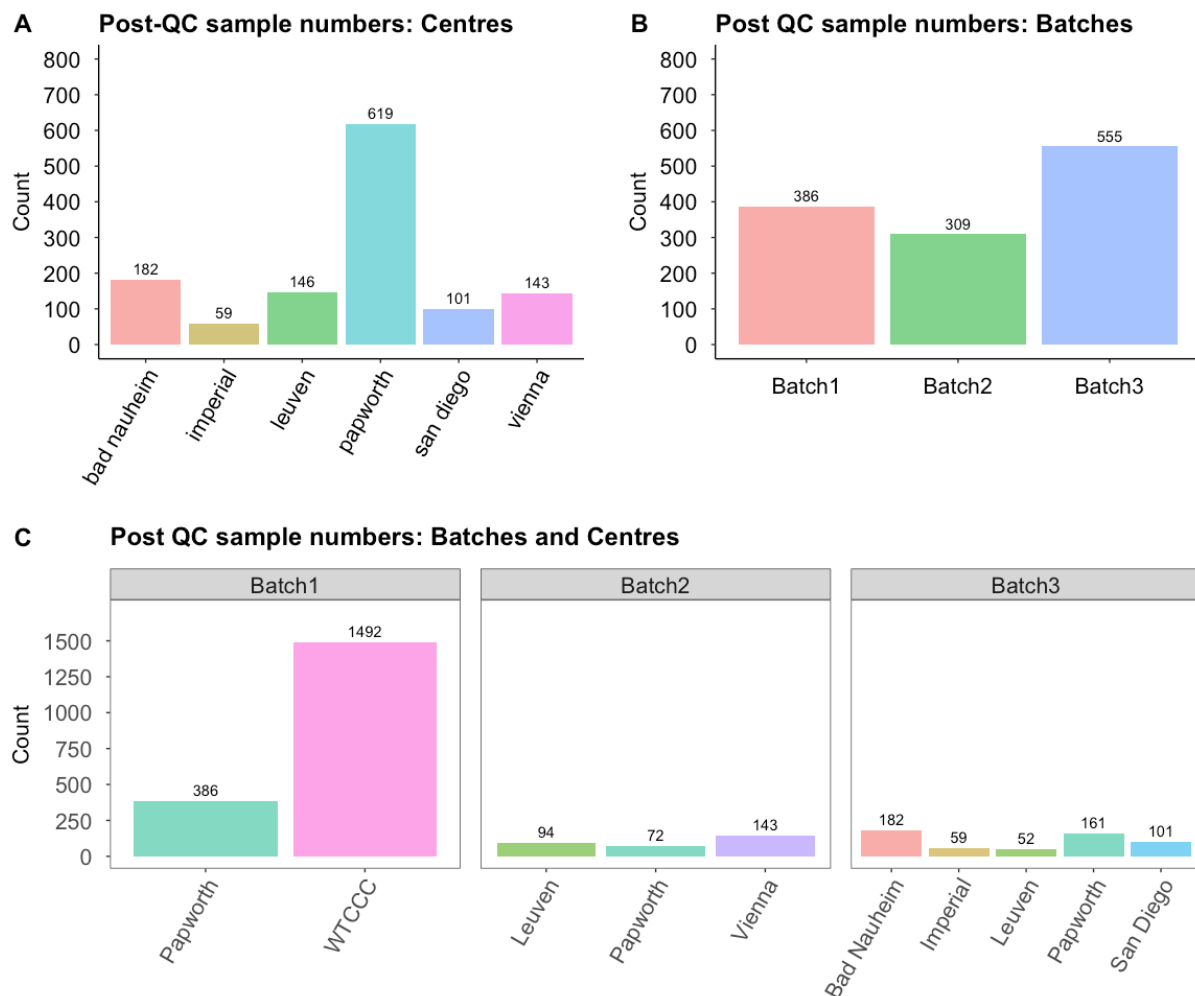


Figure 3.7 Total sample numbers following QC exclusions

Sample numbers for CTEPH cases by **A** centre and **B** batch. **C** Sample numbers for CTEPH cases and healthy controls by batch and centre. n displayed at the top of each count bar.

3.2.2.3 SNP exclusions: GWAS quality control

3.2.2.3.1 SNP exclusions: overview

A total of 31,344 (3% of total), 27,692 (3%) and 35,648 (4%) SNPs were excluded from batches1, 2 and 3 respectively during quality control steps prior to genetic

imputation ([Table 3.1](#)). SNPs were excluded on the basis of thresholds and criteria defined in [Section 2.1.4.4](#) for micro-array intensity clustering and genotype calling quality, genotype missingness and deviation from Hardy-Weinberg equilibrium. A minor allele frequency threshold was not applied until post-imputation.

	Batch1	Batch2	Batch3
Starting SNPs	964193	964193	964193
SNPs excluded	31344	27692	35648
SNPs after QC removals	932849	936501	928545

Table 3.1 Total number of SNP exclusions per batch prior to imputation

3.2.2.3.2 Micro-array clustering, genotype calling and exclusions

After the initial micro-array clustering and genotype calling using GenomeStudio(252), there were a disproportionate number of samples that were removed by applying clustering quality scores (GenTrain score < 0.7 and clustering separation score < 0.5) between batches ([Table 3.2](#)). There were 30,289 and 35,648 removed from batches 1 and 3 respectively however, 468,806 were removed from batch 2. Failure to apply clustering quality scores resulted in false positive associations ([Figure 3.8](#)). Manual inspection of micro-array data intensity plots for isolated significant SNPs confirmed that many were due to poor quality clustering ([Figure 3.9](#)).

	Batch1	Batch2	Batch3
Starting SNPs	964193	964193	964193
SNPs excluded	30289	468806	35648
GeneTrain score < 0.7			
Clustering separation score < 0.5			
SNPs after clustering QC removal	933904	495387	92854

Table 3.2 SNP exclusions from micro-array clustering quality thresholds that were applied without a re-clustering step

The number of SNPs that were excluded when clustering quality thresholds were applied after micro-array clustering and genotype calling *without* the sample missingness exclusions and re-clustering steps.

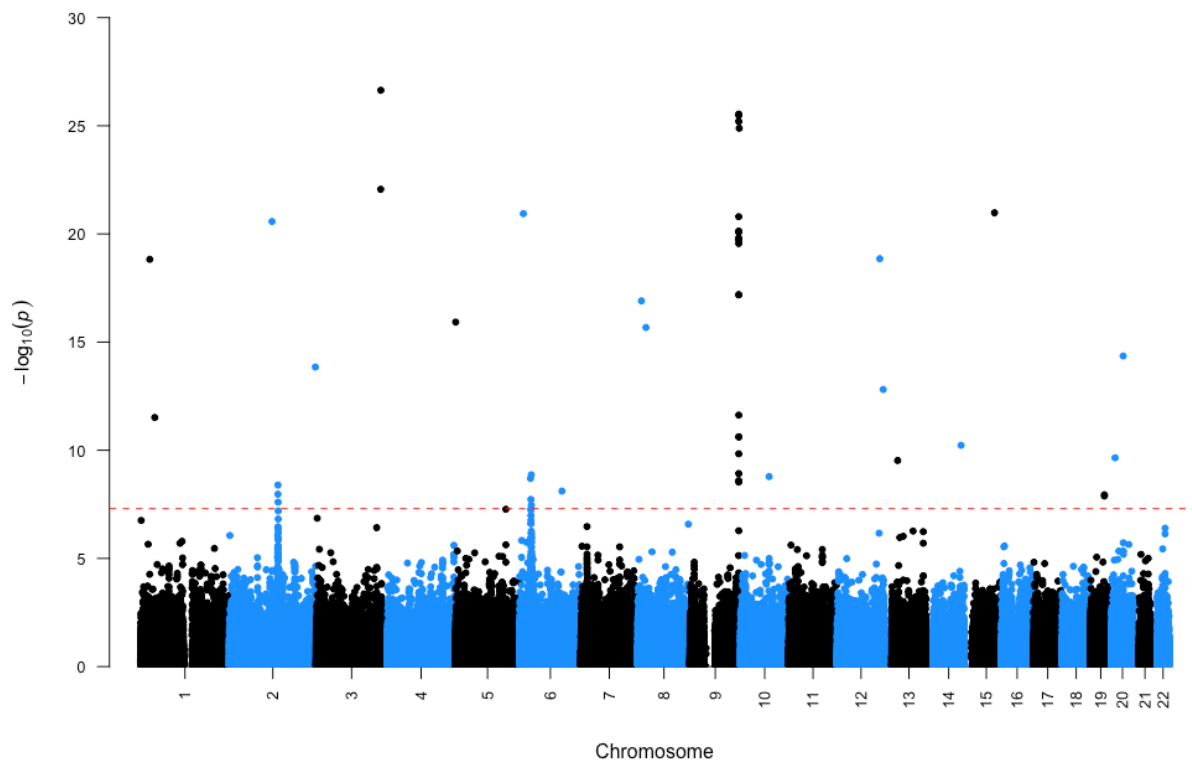


Figure 3.8 Manhattan plot of all associations including incorrect genotype clustering and calling

This was an interim analysis of batch1 and batch2 data (CTEPH cases = 900, healthy controls = 1495) with similar quality control steps for samples and SNPs as described in the methods. An additive model of association was applied using logistic regression and adjusted for 1 principal component prior to genetic imputation. Manual inspection of micro-array intensity plots for isolated significant SNPs confirmed that many were due to poor quality micro-array intensity clustering.

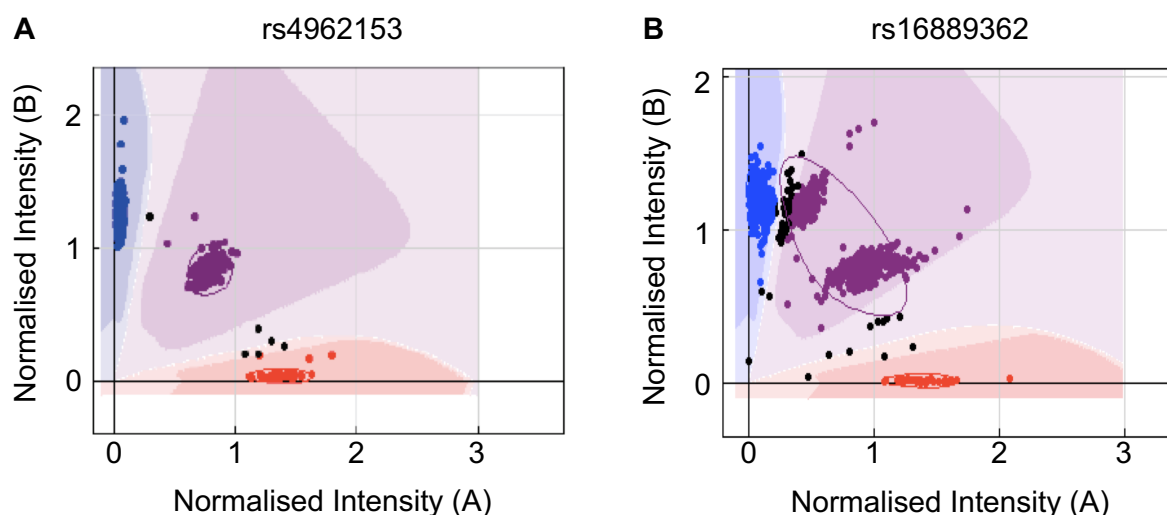


Figure 3.9 Micro-array intensity clustering and false positive associations

Micro-array intensity clustering for two biallelic SNPs with **A** good and **B** poor clustering, which fails to adequately separate the AB and BB genotypes. Intensities for Allele A and B are plotted for each SNP, which generates 3 genotypes (coloured blue, purple and red). A GenomeStudio algorithm was used for micro-array clustering and genotype calling. SNPs with poor quality clustering that are not adequately removed by quality control steps can result in false positive associations.

High missingness in a few individual samples has the potential to skew the SNP genotype calling algorithms. To address the discordant number of SNPs removed from each batch when *only* a clustering quality score was applied, the micro-array intensity clustering and genotype calling methodology was refined. First, micro-array clustering and genotype calling were performed, then samples with a missingness of greater than 1% were removed, followed by genotype re-clustering (cases and controls) and finally, SNPs were excluded using genotype clustering quality scores. This resulted in a marked improvement in batch2 SNP exclusions (n=26,198) ([Table 3.3](#)) and similar exclusions from batches 1 (n=28,124) and 3 (n=33,212) respectively. Applying a more stringent sample missingness threshold of 1% resulted in increased sample exclusions (1% sample missingness threshold: batch1=20, batch2=41, batch3=15 vs. 5% sample missingness threshold: batch1=6, batch2=15, batch3=5). This was acceptable given the marked improvement in SNP retention.

3.2.2.3.3 SNP genotype missingness, deviations from Hardy-Weinberg distribution and pre-imputation minor allele frequency

Low quality individual SNPs were excluded if their genotype missingness was greater than 1% (n: batch1=1746, batch2=1283, batch3=2180) (**Figure 3.10**). SNPs with a differential missingness ($p < 1 \times 10^{-5}$) between cases and controls were also excluded. This was only applied to batch 1 (n=1113 SNPs) as this was the only batch containing both cases and controls. As batches were subsequently merged on intersecting SNPs, then these SNP removals applied to all batches.

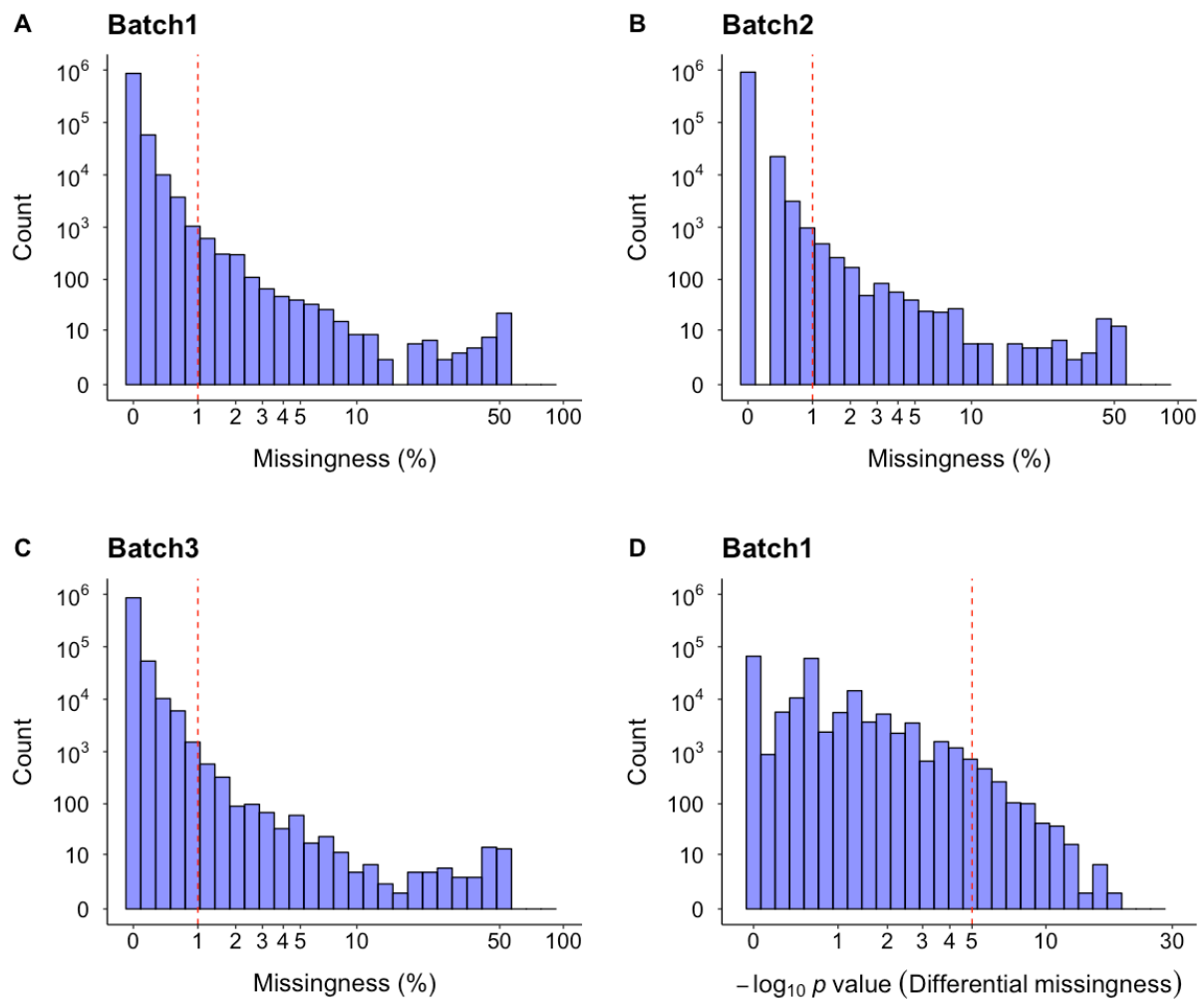


Figure 3.10 Genotype missingness SNP exclusions

The red dotted vertical line in **A**, **B** and **C** represents a missingness of 1%. The red dotted vertical line in **D** presents a p -value of 1×10^{-5} for differential missingness between CTEPH cases and healthy controls. The axes are plotted on a log1p (e.g. $\log(x+1)$) scale to allow visualisation of zero values for **A-D**. P -values have been transformed to $-\log_{10}$ values in **D**.

If SNPs are in Hardy-Weinberg equilibrium their allele and genotype frequencies can be estimated between generations. Deviation from HWE can occur due to genotyping errors, population stratification or a genuine disease association.(213) There were only a modest number of SNPs excluded due to HWE (batch1=1094, batch2=291, batch3=425) (**Figure 3.11**) potentially due to the stringent genotype clustering exclusion QC step.

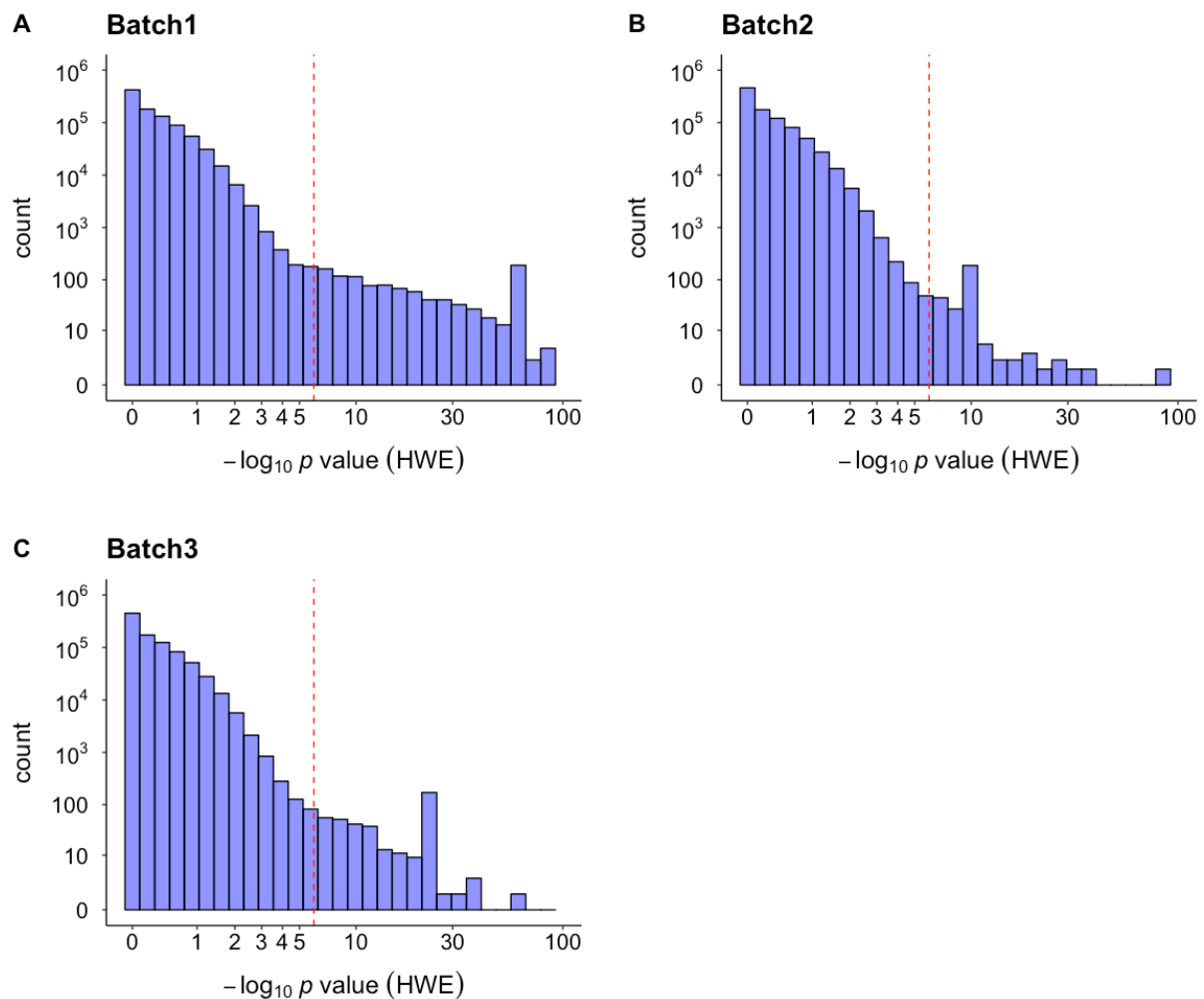


Figure 3.11 SNP Hardy-Weinberg equilibrium exclusions

The red dotted vertical line in **A-C** represents a p -value of 1×10^{-6} , below which SNPs were considered to have deviated from HWE and were excluded. The axes are plotted on a $\log_{10} p$ scale to allow visualisation of zero values and p -values have been transformed to a $-\log_{10}$ value.

SNPs were not excluded due to a minor allele frequency prior to imputation. The distribution of the minor allele frequency for each batch pre-imputation is shown in [Figure 3.12](#).

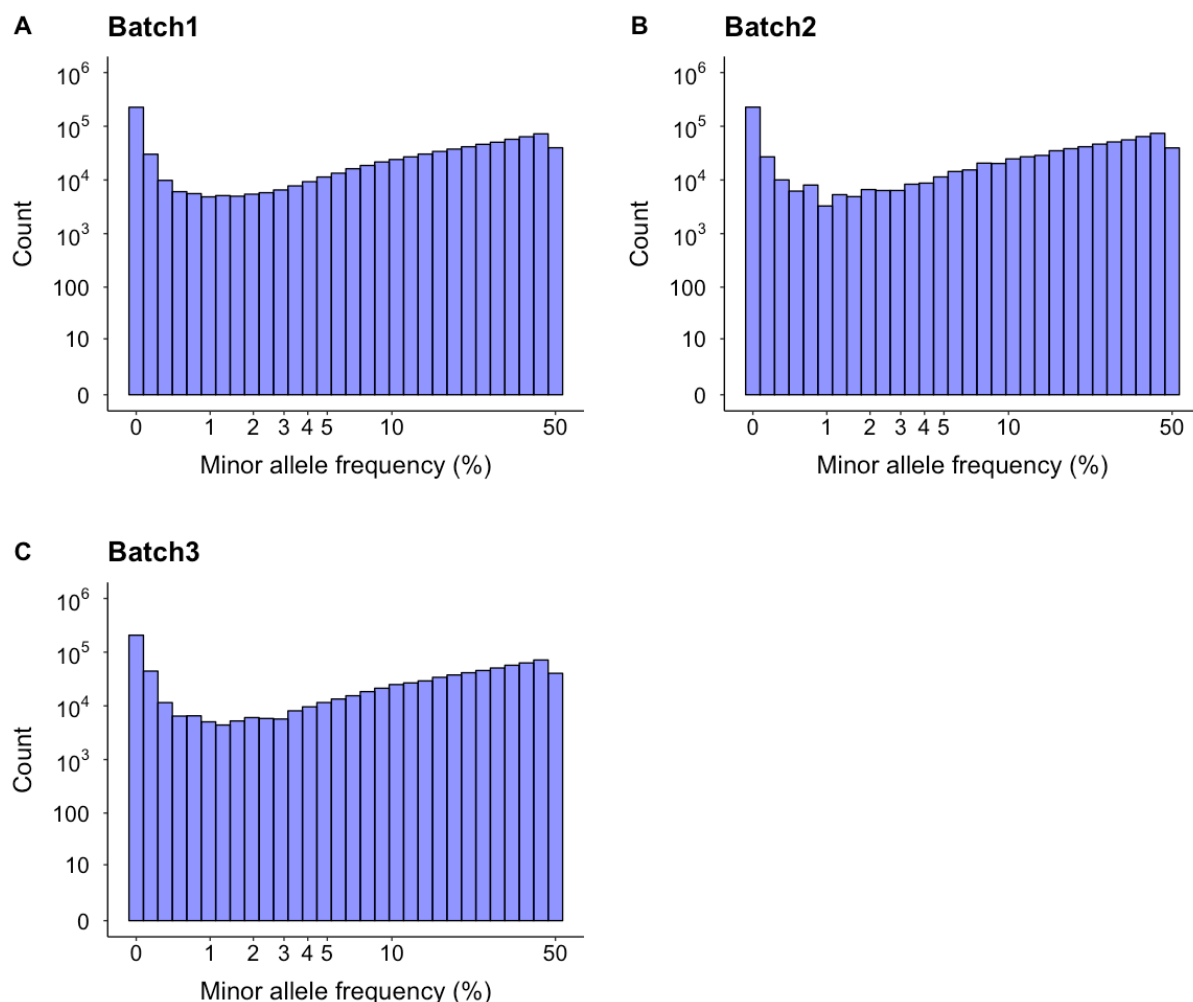


Figure 3.12 SNP minor allele frequency distribution prior to imputation

The axes are plotted on a log_{1p} scale to allow visualisation of zero values.

A summary of the QC SNP exclusions prior to imputation is shown in [Table 3.3](#). Batches 1, 2 and 3 were merged on an intersecting set of SNPs ($n=915,999$) and between-batch related samples were excluded as described in [Section 3.2.2.2](#). The data was then submitted for genetic imputation as described in the methods.

	Exclusion threshold	Batch1	Batch2	Batch3
Starting SNPs		964193	964193	964193
Genotype clustering exclusions	GenTrain < 0.7 Clustering separation < 0.5	28124	26198	33212
SNP missingness exclusions	> 1%	1746	1283	2180
SNP differential missingness exclusions	$p < 1 \times 10^{-5}$	1113	NA	NA
Divergent Hardy-Weinberg equilibrium	$p < 1 \times 10^{-6}$	1094	291	425
Total (unique) SNPs excluded		31344	27692	35648
SNPS after QC removals		932849	936501	928545

Table 3.3 SNP exclusions for each quality control step

SNPs could be excluded on more than one criterion therefore, the individual SNP exclusions do not equal the "Total (unique)" values.

3.2.2.3.4 Post GWAS QC: minor allele frequency and imputation quality

Following genetic imputation there were ~40 million variants. Imputation quality was assessed using an information (INFO) score parameter. This ranged on a scale of 0-1 from poor to high quality imputation ([Figure 3.13](#)). Variants with an INFO score below 0.5 were excluded. SNPs with low minor allele frequencies can be challenging for genotype clustering algorithms to call and are also more difficult to impute, which can result in false positive associations.(142) SNPs were excluded if they had a low minor allele frequency (<1%) ([Figure 3.14](#)).

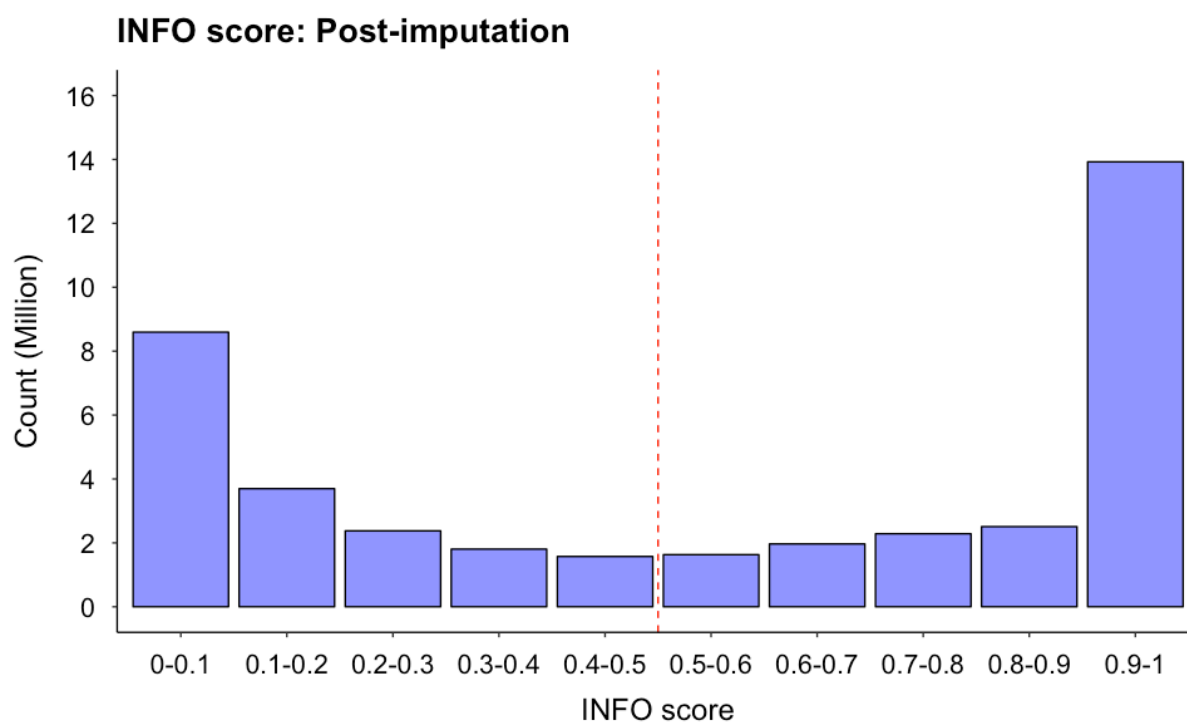


Figure 3.13 SNP imputation quality

The information (INFO) score was a parameter supplied following genetic imputation to assess the imputation quality. The dashed vertical red line represents an information score threshold of 0.5 below which, SNPs were excluded (n=18,043,895).

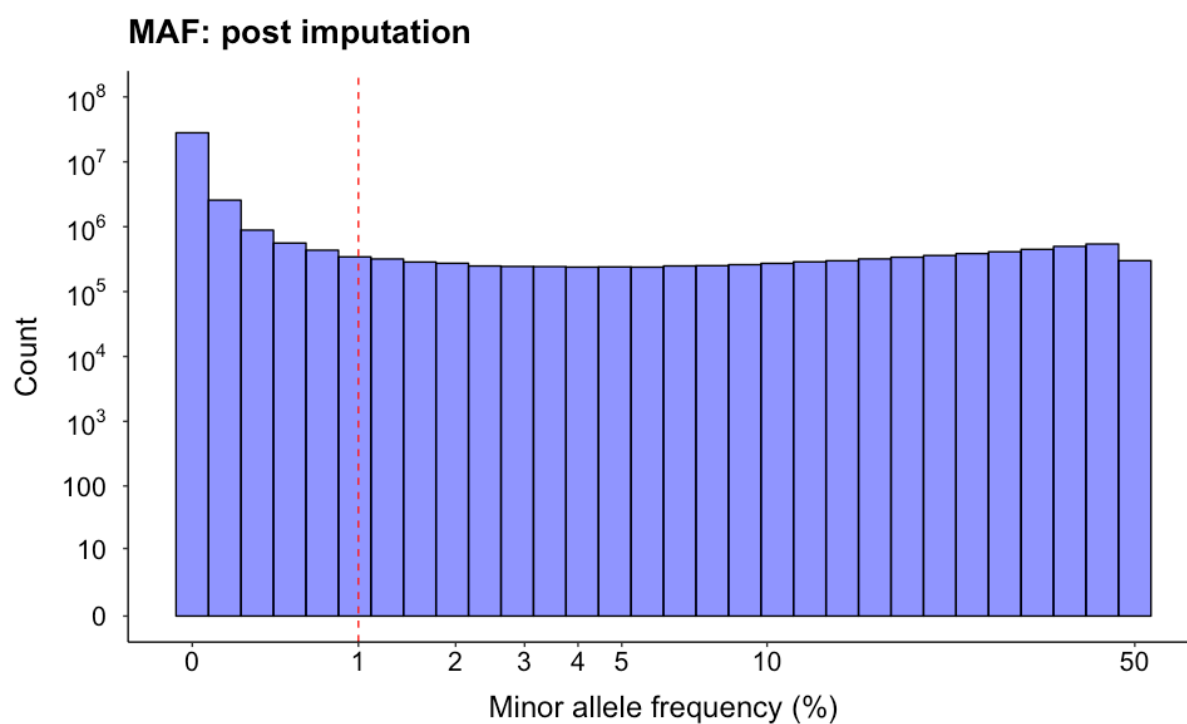


Figure 3.14 Minor allele frequency post imputation

The dashed red vertical line represents a MAF threshold of 1% below which, variants were excluded (n=32,676,650). The axes are plotted on a log₁₀p scale to allow visualisation of zero values.

After removal of SNPs with low imputation quality, low minor allele frequency and multi-allelic SNPs there were 7,675,738 remaining for association testing.

3.2.2.4 Residual population structure

Residual population structure was assessed after QC exclusions and prior to imputation for each batch and after merging batches 1, 2 and 3 on a common intersecting SNP set. PCAs were performed and principal components were visualised using up to 20 eigenvector pairs to detect outlying clusters ([Figures 3.15](#) and [3.16](#); the first 5 eigenvector pairs are shown). These principal components were used for to adjust for residual population structure in subsequent association testing.

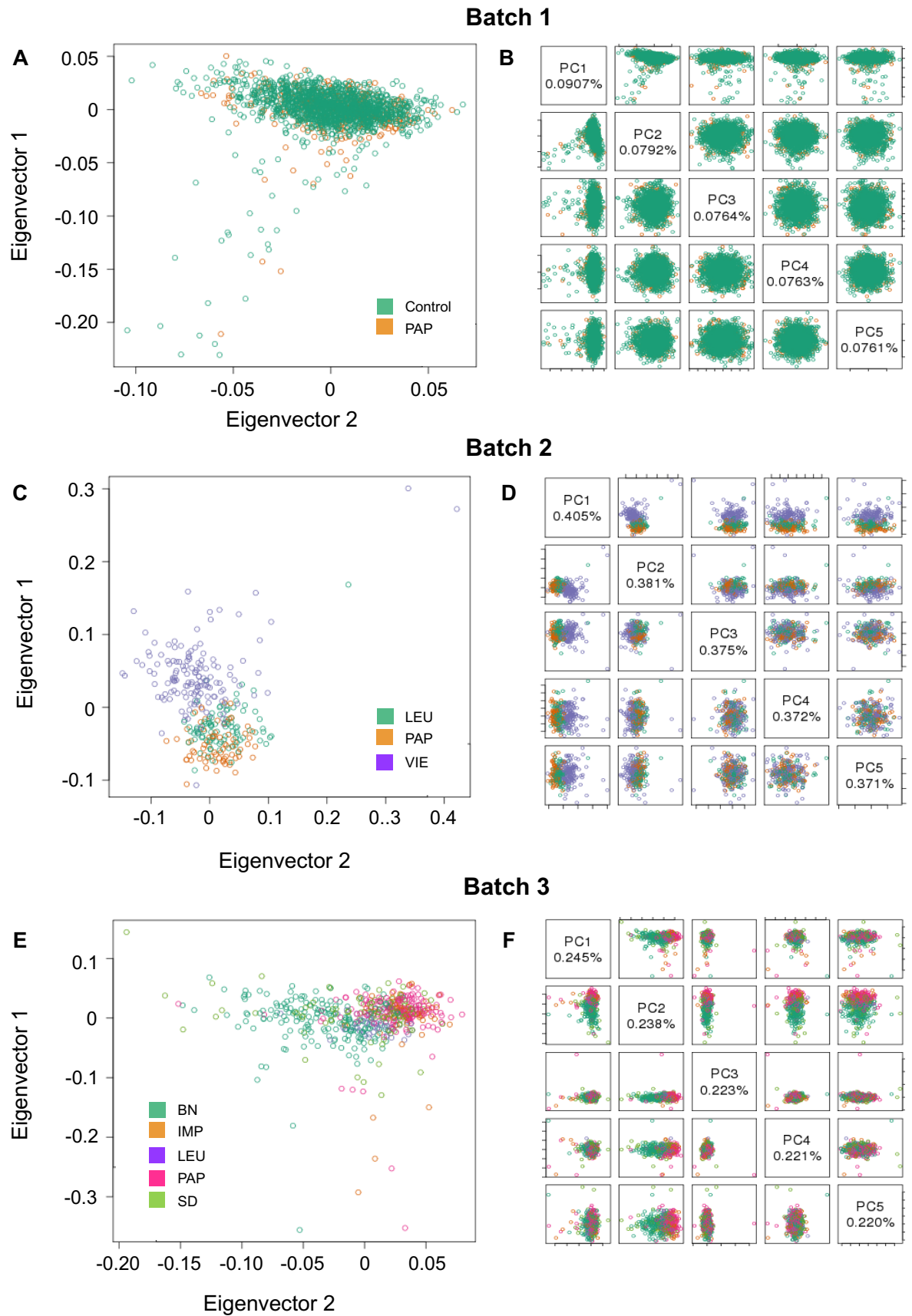


Figure 3.15 Principal component analysis to detect residual population structure for each batch

PCAs were performed on individual batches *prior* to merging on a common SNP set. The first two eigenvectors are displayed in **A**, **C** and **E**, and the first 5 pairs of eigenvectors are shown in **B**, **D** and **F** including the percentage of variation that is explained by the pair. Up to 20 eigenvector pairs were visualised (not shown). Colours used to represent centres vary between plots.

BN (Bad Nauheim), IMP (Imperial), LEU (Leuven), PAP (Papworth), SD (San Diego), VIE (Vienna), WTCCC (Wellcome Trust Case Control Consortium).

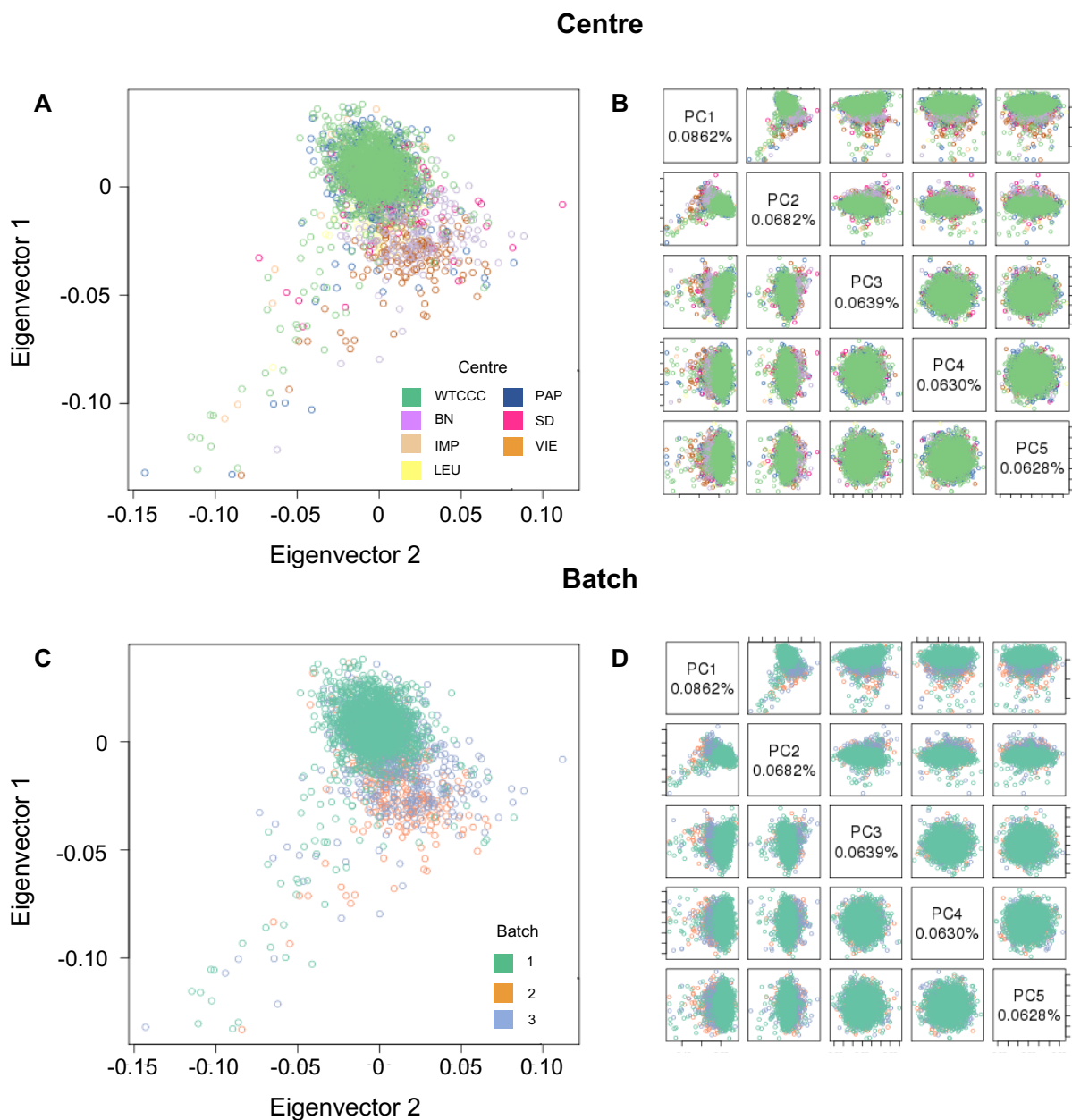


Figure 3.16 Principal component analysis to detect residual population structure for combined batches

PCAs were performed for all CTEPH cases and healthy controls following QC exclusions and merging batches, but prior to genetic imputation. PCA plots have been coloured by **A, B** centre and **C, D** batch. The first two eigenvectors are displayed in **A, C**, and the first 5 pairs of eigenvectors are shown in **B, D** including the percentage of variation that is explained by the pair. Up to 20 eigenvector pairs were visualised (not shown).

BN (Bad Nauheim), IMP (Imperial), LEU (Leuven), PAP (Papworth), SD (San Diego), VIE (Vienna), WTCCC (Wellcome Trust Case Control Consortium).

3.2.2.5 Study participant characteristics post-QC

The baseline characteristics of the case-control groups following QC exclusions are summarised in [Table 3.4](#). There was no difference in sex however, CTEPH cases were older than healthy controls (median \pm IQR: 65 \pm 22 vs. 45 \pm 18; $p < 0.001$). It was not necessary to adjust for age in the GWAS statistical association testing as the prevalence of CTEPH in the general population ($< 1/30,000$) is low and any confounding due to the presence of CTEPH in the healthy control group (unidentified or new incident cases with increasing age) was unlikely.

	Healthy Controls	CTEPH cases	<i>p</i>
Study participants, n	1492	1250	
Sex: Female, n (%)	795 (52)	712 (49)	0.164
Age, median \pm IQR	45 \pm 18	65 \pm 22	< 0.001

Table 3.4 Baseline characteristics for the case-control groups included in association testing

Group differences in sex and age assessed by Chi-squared test and Mann-Whitney U test respectively.

3.2.3 GWAS statistical association testing

Following genetic imputation, the data for the discovery and validation cohorts were combined in one dataset. A joint analysis was initially performed prior to dividing the data into discovery and validation cohorts. This also enabled selection of the most appropriate number of principal components to adjust for residual population structure that were then applied to the discovery and validation groups. The composition of the discovery and validation cohorts was determined *a priori*. The discovery cohort was comprised of UK centres and the remaining European and US centres made up the validation cohort.

3.2.3.1 Joint analysis: discovery and validation cohorts combined

3.2.3.1.1 Association testing without covariates

Imputed genotype dosages were used to test for an association between the CTEPH and healthy control groups. Logistic regression assuming an additive genetic model was applied to each SNP marker. When no covariates were used to adjust the models there are at least 4 associated loci, however the genomic inflation factor is elevated ($\lambda = 1.22$) ([Figure 3.17](#)). The inflation factor is estimated by comparing the median of observed and expected test statistics.⁽¹⁴⁵⁾ Genomic inflation values above 1 can indicate population stratification or genotyping errors.⁽¹⁴⁵⁾ When the data is divided into discovery and validation cohorts ([Sections 3.2.3.2](#) and [3.2.3.3](#)) and association testing repeated without additional covariates, the genomic inflation is higher in the validation cohort ($\lambda = 1.03$ vs. 1.47) ([Figure 3.18](#)). As samples comprising the discovery and validation groups were genotyped across batches, the elevated genomic inflation values are a likely consequence of population stratification. This can result in false positive associations due to differences in ancestry rather than a genuine disease association.⁽²¹³⁾ To adjust for confounding from population stratification, eigenvectors from the pre-imputation PCAs ([Section 3.2.2.4](#)) were included as covariates in the logistic regression models.

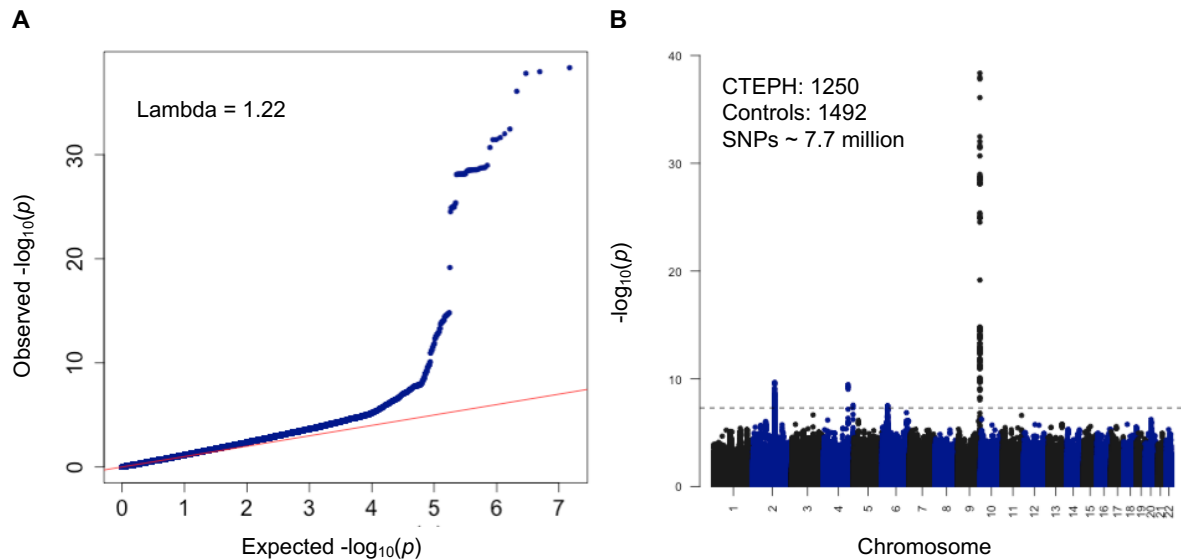


Figure 3.17 Case-control association testing without additional covariates: joint analysis

Analysis of 1250 CTEPH cases (discovery and validation cohorts combined), 1492 healthy controls and 7,675,738 SNPs. The same healthy control group was used for the discovery and validation cohorts. Statistical testing of individual SNPs using allelic dosage (range 0-2) for an association with CTEPH diagnosis was performed using logistic regression assuming an additive genetic model without additional covariates. A p -value of $<5 \times 10^{-8}$ was considered genome-wide significant (dotted grey line **B**). Genomic inflation factor (λ)=1.22.

A Quantile-quantile (QQ) plot of the observed and expected p -values.

B Manhattan plot of p -values plotted against genomic position.

P -values are transformed to a $-\log_{10}$ scale.

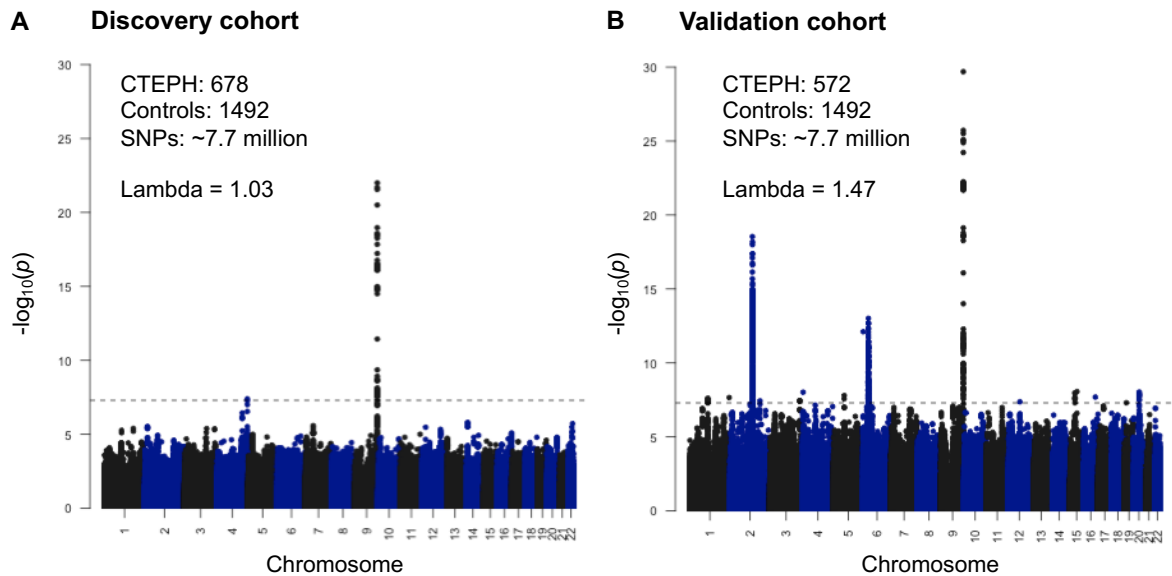


Figure 3.18 Case-control association testing without covariates: discovery and validation cohorts

Manhattan plots of **A** Discovery and **B** Validation cohorts.

Statistical testing without additional covariates was performed as described in [Figure 3.17](#). The genomic inflation factor (lambda) for the discovery and validation cohorts are 1.03 and 1.47 respectively.

3.2.3.1.2 Association testing adjusted for population stratification

The first 2 eigenvectors were most significantly associated ($p=1.48 \times 10^{-57}$ and $p=2.35 \times 10^{-30}$) with the case-control group when logistic regression was performed ([Table 3.5](#)). As the fifth eigenvector was nominally associated, the first 5 ancestry informative principal components were used to adjust for residual population stratification. This markedly improved the genomic inflation (lambda = 1.04) and confirmed that the elevation was due to residual population substructure ([Figure 3.19](#)). As other eigenvectors were nominally associated (EV11 and EV16) with case-control status, up to 20 were included as covariates in GWAS association testing. These did not improve the genomic inflation further (lambda: EV1-10 = 1.05, EV1-20 = 1.05) and therefore, only the first 5 ancestry informative principal components were used for subsequent association testing.

	β	SE	p
EV1	-43.0	2.69	1.48e-57
EV2	30.0	2.62	2.35e-30
EV3	-1.20	2.32	0.604
EV4	0.884	2.22	0.690
EV5	3.82	2.21	0.084
EV6	1.32	2.21	0.550
EV7	-0.376	2.20	0.864
EV8	0.705	2.20	0.748
EV9	2.71	2.21	0.220
EV10	-1.86	2.19	0.396
EV11	-6.51	2.22	0.003
EV12	0.306	2.20	0.889
EV13	2.78	2.21	0.209
EV14	0.74	2.20	0.737
EV15	-0.0785	2.20	0.972
EV16	-3.98	2.20	0.071
EV17	-0.0619	2.20	0.978
EV18	0.237	2.19	0.914
EV19	2.79	2.20	0.205
EV20	3.40	2.20	0.122

Table 3.5 Ancestry informative eigenvectors and case-control status

Logistic regression of eigenvectors (EV) 1-20 on case-control status (case/control group \sim EV1 + EV2 + ... + EV20). The most significant eigenvectors are EV1 and EV2. β (beta coefficient), SE (standard error), p (p -value).

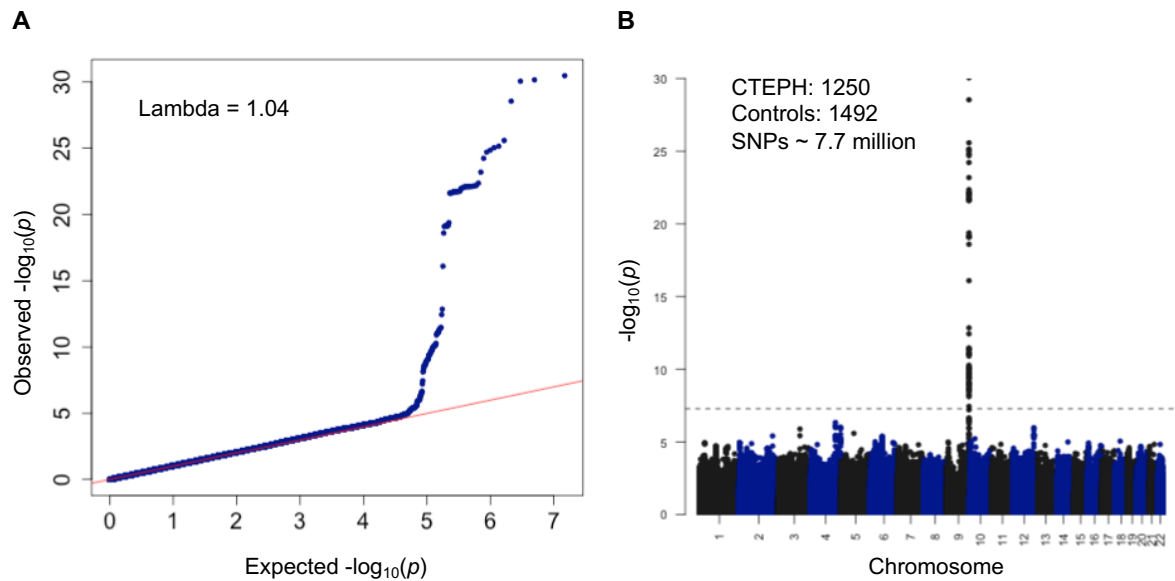


Figure 3.19 Case-control association testing with 5 ancestry informative principal components: joint analysis

Case-control association testing of 1250 CTEPH cases and 1492 controls using allelic dosages and assuming an additive genetic model adjusted for the first 5 ancestry informative principal components.

A QQ plot and **B** Manhattan plot

Genomic inflation factor (λ)=1.04

There were 87 significantly associated SNPs ($p < 5 \times 10^{-8}$) (Table 3.6) that were all located in chromosome 9. The lead SNP (rs2519093, OR (95% CI) = 2.4 (2.3-2.5), $p = 3.42 \times 10^{-31}$) is an intronic variant in the *ABO* gene. Rs2519093 was well imputed (INFO=0.951) with an effect allele (T) frequency of 0.165 in healthy controls and 0.316 in CTEPH cases. This compared with effect allele frequencies in reference populations of 0.182 (European (non-Finnish) in 1000 genomes phase 3) and 0.198 (European (non-Finnish) in gnomAD).(240)

rsID	CHR:POS_EA/NEA	GENE	FUNC	EAF_A	EAF_U	EAF_REF	INFO	OR (95% CI)	p
rs2519093	9:136141870_T/C	ABO	intronic	0.316	0.165	0.182	0.951	2.4 (2.25-2.55)	3.42e-31
rs532436	9:136149830_A/G	ABO	intronic	0.316	0.165	0.185	0.965	2.39 (2.24-2.53)	7.01e-31
rs507666	9:136149399_A/G	ABO	intronic	0.316	0.166	0.185	0.965	2.38 (2.23-2.53)	8.89e-31
rs635634*	9:136155000_T/C	ABO	intergenic	0.321	0.174	0.185	0.977	2.31 (2.16-2.46)	2.87e-29
rs600038	9:136151806_C/T	ABO	intergenic	0.331	0.189	0.218	0.965	2.16 (2.02-2.3)	2.61e-26
rs651007*	9:136153875_T/C	ABO	intergenic	0.330	0.189	0.215	0.971	2.13 (1.99-2.27)	7.1e-26
rs579459*	9:136154168_C/T	ABO	intergenic	0.330	0.189	0.215	1.000	2.12 (1.98-2.26)	9.08e-26
rs649129	9:136154304_T/C	ABO	intergenic	0.330	0.189	0.215	0.988	2.12 (1.98-2.27)	1.44e-25
rs495828	9:136154867_T/G	ABO	intergenic	0.330	0.189	0.215	0.964	2.13 (1.99-2.27)	2e-25
rs550057	9:136146597_T/C	ABO	intronic	0.386	0.241	0.280	0.972	2.03 (1.9-2.17)	5.88e-25
rs9411378	9:136145425_A/C	ABO	intronic	0.380	0.236	0.290	0.887	2.09 (1.95-2.23)	6.5e-24
rs529565	9:136149500_C/T	ABO	intronic	0.463	0.313	0.345	0.976	1.9 (1.77-2.02)	4.42e-23
rs505922*	9:136149229_C/T	ABO	intronic	0.463	0.313	0.345	1.000	1.88 (1.75-2)	6.75e-23
rs582094	9:136145484_T/A	ABO	intronic	0.463	0.315	0.350	0.990	1.89 (1.76-2.01)	7.08e-23
rs2769071	9:136145974_G/A	ABO	intronic	0.463	0.315	0.350	0.971	1.9 (1.77-2.03)	7.81e-23
rs582118	9:136145471_G/A	ABO	intronic	0.463	0.315	0.350	0.991	1.88 (1.76-2.01)	7.89e-23
rs676996	9:136146077_G/T	ABO	intronic	0.463	0.315	0.353	0.991	1.88 (1.76-2.01)	8.1e-23
rs597988	9:136144284_A/T	ABO	intronic	0.463	0.315	0.348	0.992	1.88 (1.76-2.01)	8.11e-23
rs677355	9:136146046_A/G	ABO	intronic	0.463	0.315	0.353	0.971	1.9 (1.77-2.03)	8.12e-23
rs492488	9:136144960_A/G	ABO	intronic	0.463	0.315	0.350	0.989	1.88 (1.76-2.01)	8.74e-23
rs676457	9:136146227_T/A	ABO	intronic	0.463	0.315	0.350	0.991	1.88 (1.75-2.01)	1.05e-22
rs687289*	9:136137106_A/G	ABO	intronic	0.463	0.315	0.353	0.989	1.88 (1.76-2.01)	1.1e-22

rs493246	9:136144994_A/G	ABO	intronic	0.463	0.315	0.350	0.999	1.87 (1.74-1.99)	1.79e-22
rs495203	9:136145240_T/C	ABO	intronic	0.463	0.315	0.350	0.999	1.86 (1.74-1.99)	1.8e-22
rs514659*	9:136142203_C/A	ABO	intronic	0.463	0.315	0.350	1.000	1.86 (1.74-1.99)	1.92e-22
rs8176663	9:136144427_C/T	ABO	intronic	0.463	0.315	0.350	1.000	1.86 (1.74-1.99)	1.93e-22
rs491626	9:136144873_T/C	ABO	intronic	0.463	0.315	0.350	1.000	1.86 (1.74-1.99)	1.93e-22
rs545971	9:136143372_T/C	ABO	intronic	0.463	0.315	0.350	1.000	1.86 (1.74-1.99)	1.94e-22
rs612169	9:136143442_G/A	ABO	intronic	0.463	0.315	0.348	1.000	1.86 (1.74-1.99)	1.94e-22
rs687621*	9:136137065_G/A	ABO	intronic	0.463	0.315	0.350	1.000	1.86 (1.74-1.99)	2.27e-22
rs527210	9:136146431_T/C	ABO	intronic	0.462	0.315	0.350	0.977	1.88 (1.75-2.01)	2.33e-22
rs674302	9:136146664_A/T	ABO	intronic	0.463	0.315	0.350	0.999	1.86 (1.74-1.99)	2.41e-22
rs554833	9:136147160_T/C	ABO	intronic	0.463	0.315	0.350	0.997	1.86 (1.74-1.99)	2.46e-22
rs494242	9:136145118_T/C	ABO	intronic	0.478	0.339	0.377	0.985	1.79 (1.67-1.91)	4.19e-20
rs644234	9:136142217_G/T	ABO	intronic	0.478	0.339	0.377	0.994	1.78 (1.65-1.9)	7.08e-20
rs8176645	9:136149098_A/T	ABO	intronic	0.463	0.322	0.375	0.735	2.02 (1.86-2.17)	7.23e-20
rs613534	9:136143120_G/A	ABO	intronic	0.478	0.339	0.377	0.995	1.78 (1.65-1.9)	7.42e-20
rs543968	9:136143121_C/T	ABO	intronic	0.478	0.339	0.377	0.995	1.78 (1.65-1.9)	7.42e-20
rs544873	9:136143212_A/G	ABO	intronic	0.478	0.339	0.377	0.995	1.78 (1.65-1.9)	7.42e-20
rs643434	9:136142355_A/G	ABO	intronic	0.478	0.339	0.377	0.996	1.77 (1.65-1.9)	8.2e-20
rs657152*	9:136139265_A/C	ABO	intronic	0.478	0.339	0.377	1.000	1.76 (1.63-1.88)	2.54e-19
rs11244061	9:136153981_T/C	ABO	intergenic	0.187	0.102	0.120	0.947	2.1 (1.93-2.28)	7.95e-17
rs11244084	9:136191010_T/C	LCN1P2	intergenic	0.165	0.091	0.075	0.850	2.03 (1.84-2.22)	1.38e-13
rs142956930	9:136143330_G/A	ABO	intronic	0.131	0.066	0.017	0.513	3.31 (2.99-3.63)	3.57e-13
rs8176681	9:136139754_T/C	ABO	intronic	0.654	0.548	0.592	0.990	1.54 (1.42-1.66)	3.36e-12

rs2073827	9:136137133_G/C	<i>ABO</i>	intronic	0.653	0.548	0.592	0.989	1.54 (1.42-1.66)	3.58e-12
rs2073828	9:136137140_G/A	<i>ABO</i>	intronic	0.653	0.548	0.595	0.989	1.54 (1.41-1.66)	4.34e-12
rs559723	9:136150484_A/G	<i>ABO</i>	intronic	0.382	0.490	0.498	0.988	1.53 (1.41-1.65)	4.67e-12
rs616154	9:136150466_C/T	<i>ABO</i>	intronic	0.382	0.491	0.500	0.976	1.53 (1.41-1.65)	7.16e-12
rs630014*	9:136149722_A/G	<i>ABO</i>	intronic	0.381	0.489	0.495	1.000	1.52 (1.4-1.64)	8.23e-12
rs630510	9:136149581_A/G	<i>ABO</i>	intronic	0.381	0.489	0.495	0.998	1.52 (1.4-1.64)	8.39e-12
rs8176690	9:136138317_A/G	<i>ABO</i>	intronic	0.653	0.548	0.592	0.980	1.53 (1.41-1.65)	8.97e-12
rs2073826	9:136136963_G/T	<i>ABO</i>	intronic	0.652	0.548	0.595	0.977	1.52 (1.4-1.65)	1.22e-11
rs8176715	9:136133148_T/C	<i>ABO</i>	intronic	0.379	0.481	0.410	0.917	1.52 (1.4-1.65)	5.19e-11
rs8176668	9:136144059_A/T	<i>ABO</i>	intronic	0.669	0.572	0.617	0.977	1.5 (1.38-1.63)	7.19e-11
rs7873635	9:136132012_T/C	<i>ABO</i>	intronic	0.333	0.428	0.383	0.831	1.58 (1.45-1.72)	7.56e-11
rs7046674	9:136147012_C/T	<i>ABO</i>	intronic	0.669	0.571	0.620	0.991	1.5 (1.38-1.62)	7.87e-11
rs8176649	9:136147295_G/A	<i>ABO</i>	intronic	0.669	0.571	0.620	0.992	1.5 (1.38-1.62)	7.95e-11
rs7036642	9:136144626_G/A	<i>ABO</i>	intronic	0.669	0.572	0.620	0.985	1.5 (1.38-1.62)	9.6e-11
rs3124761	9:136339755_C/T	<i>SLC2A6</i>	intronic	0.799	0.877	0.838	0.876	1.76 (1.59-1.93)	9.77e-11
rs3094379	9:136334910_C/T	<i>CACFD1</i>	UTR3	0.798	0.876	0.838	0.887	1.75 (1.58-1.92)	1.25e-10
rs8176691	9:136138229_C/T	<i>ABO</i>	intronic	0.668	0.572	0.620	0.989	1.49 (1.37-1.61)	1.78e-10
rs8176682*	9:136139297_C/T	<i>ABO</i>	intronic	0.668	0.572	0.620	0.990	1.49 (1.37-1.61)	1.86e-10
rs3124764	9:136329954_C/T	<i>CACFD1</i>	intronic	0.800	0.877	0.840	0.892	1.74 (1.57-1.91)	1.91e-10
rs3124765	9:136328657_C/T	<i>CACFD1</i>	exonic	0.800	0.877	0.892	0.892	1.74 (1.57-1.91)	2.13e-10
rs4962153*	9:136323754_G/A	<i>ADAMTS13</i>	intronic	0.800	0.877	0.838	0.900	1.72 (1.55-1.89)	3.01e-10
rs3124766	9:136316942_A/G	<i>ADAMTS13</i>	intronic	0.200	0.124	0.162	0.906	1.71 (1.54-1.88)	3.4e-10
rs28645493*	9:136305738_G/C	<i>ADAMTS13</i>	intronic	0.140	0.076	0.098	1.000	1.86 (1.66-2.05)	3.44e-10

rs739468	9:136326248_G/T	<i>CACFD1</i>	intronic	0.799	0.875	0.838	0.897	1.71 (1.54-1.88)	4.07e-10
rs8176702	9:136136146_G/A	<i>ABO</i>	intronic	0.667	0.572	0.620	0.976	1.48 (1.36-1.6)	4.19e-10
rs4962040	9:136133531_A/G	<i>ABO</i>	intronic	0.666	0.571	0.617	0.973	1.47 (1.35-1.6)	6.55e-10
rs500499	9:136148648_G/C	<i>ABO</i>	intronic	0.362	0.454	0.432	0.959	1.47 (1.35-1.6)	8.6e-10
rs500498	9:136148647_T/C	<i>ABO</i>	intronic	0.362	0.454	0.432	0.959	1.47 (1.35-1.6)	8.71e-10
rs36222279	9:136315974_C/G	<i>ADAMTS13</i>	intronic	0.139	0.076	0.098	0.926	1.86 (1.66-2.06)	9.92e-10
rs476410	9:136148368_G/C	<i>ABO</i>	intronic	0.362	0.454	0.432	0.961	1.47 (1.35-1.59)	1.01e-09
rs41302667	9:136330428_A/G	<i>CACFD1</i>	intronic	0.138	0.076	0.098	0.911	1.87 (1.67-2.07)	1.18e-09
rs645982	9:136148409_A/G	<i>ABO</i>	intronic	0.362	0.453	0.432	0.961	1.47 (1.34-1.59)	1.24e-09
rs28602660	9:136312071_A/G	<i>ADAMTS13</i>	intronic	0.140	0.077	0.100	0.938	1.85 (1.65-2.04)	1.41e-09
rs475419	9:136148231_C/T	<i>ABO</i>	intronic	0.362	0.453	0.432	0.979	1.45 (1.33-1.58)	2.17e-09
rs660340	9:136147553_A/G	<i>ABO</i>	intronic	0.362	0.453	0.432	0.979	1.45 (1.33-1.58)	2.2e-09
rs581107	9:136147702_C/T	<i>ABO</i>	intronic	0.362	0.453	0.432	0.979	1.45 (1.33-1.57)	2.24e-09
rs28446901	9:136308796_G/C	<i>ADAMTS13</i>	intronic	0.234	0.149	0.180	0.920	1.62 (1.46-1.78)	2.34e-09
rs473533	9:136148035_T/C	<i>ABO</i>	intronic	0.362	0.453	0.432	0.962	1.45 (1.33-1.58)	2.88e-09
rs659104	9:136147823_T/G	<i>ABO</i>	intronic	0.362	0.453	0.432	0.970	1.45 (1.32-1.57)	4.1e-09
rs633862	9:136155444_C/T	<i>ABO</i>	intergenic	0.367	0.455	0.448	0.971	1.43 (1.31-1.55)	6e-09
rs558240*	9:136157133_A/G	<i>ABO</i>	intergenic	0.492	0.402	0.405	1.000	1.4 (1.29-1.52)	7.44e-09
rs647800	9:136148000_A/G	<i>ABO</i>	intronic	0.347	0.429	0.405	0.975	1.41 (1.29-1.54)	3.48e-08

Table 3.6 Significant SNPs in the joint analysis

Association testing for 1250 CTEPH cases, 1492 healthy controls and 7,675,738 SNPs post-imputation were included in the analysis. Statistical testing of individual SNPs using allelic dosage (range 0-2) for an association with CTEPH diagnosis was performed using logistic regression assuming an additive genetic model. 5 principal components were used to adjust for any residual population structure. A p -value of $<5 \times 10^{-8}$ was considered genome-wide significant and SNPs are ordered by significance. The Genome Reference Consortium human genome (build 37) (GRCh37) was used for genomic positions. The allele frequencies and the odds ratios are for the effect alleles. P -values are displayed using exponential notation. * SNPs with an asterisk were present on the micro-array chip pre-imputation and those without have been imputed. rsID (reference SNP cluster ID), CHR (chromosome), POS (base position), EF (effect allele), NEA (non-effect allele), GENE (nearest gene for the SNP, from ANNOVAR), FUNC (functional consequence of the SNP on the gene, from ANNOVAR), EAF_A (effect allele frequency of CTEPH patients), EAF_U (effect allele frequency of healthy controls), EAF_REF (effect allele frequency of reference, 1000 genomes phase 3 European (non-Finnish) populations), INFO (information score, imputation quality), OR (odds ratio), p (p -value).

To ensure that the disease associations for CTEPH were not being confounded by recruiting centre or genotyping batch, separate *within* case analyses that included only CTEPH patients were undertaken. Linear regression was performed with batch or centre as the dependent variable and the SNP allelic dosage as the independent variable, additionally adjusted for 5 ancestry informative principal components. This confirmed that centre and batch were not major sources of confounding ([Figure 3.20](#)).

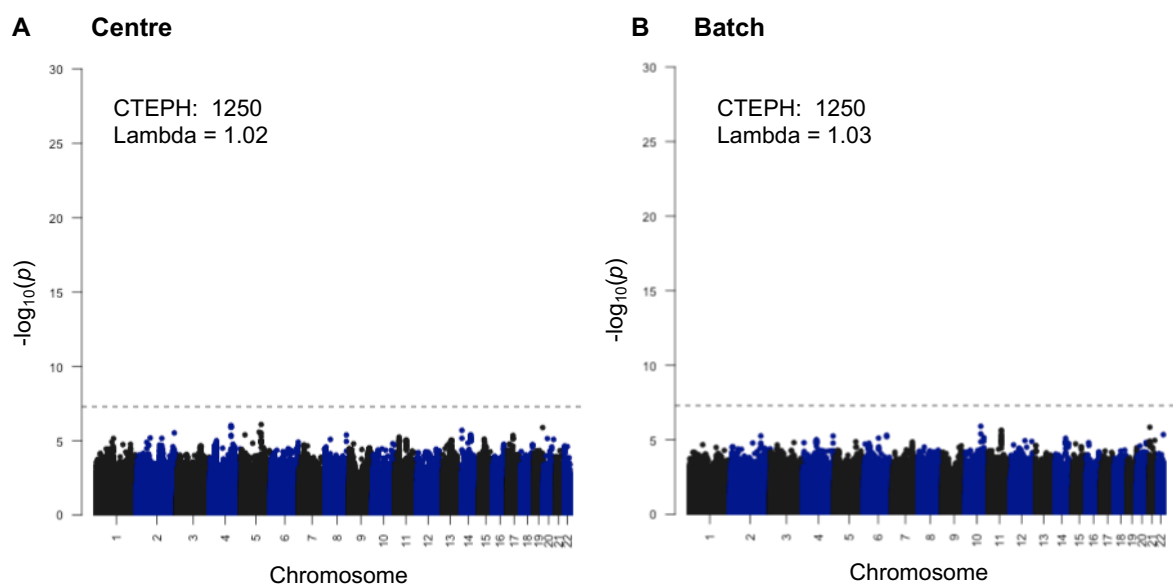


Figure 3.20 Batch and centre association testing with CTEPH

Linear regression was performed with batch or centre as the dependent variable with the CTEPH group ($n=1250$): Batch or centre \sim SNP allelic dosage + EV1 + ... + EV5.

3.2.3.1.3 Independent associations

The majority of the associated SNPs in the chromosome 9 locus were in close proximity to the *ABO* gene ([Figure 3.21](#)). There was an additional cluster of significant SNPs around the *ADAMTS13* gene that were in moderate to low linkage disequilibrium with the significant *ABO* SNPs ([Figures 3.21](#) and [3.22](#)). These *ADAMTS13* SNPs were not significant when the analysis was conditioned on the lead SNP in the *ABO* locus (rs2519093), indicating that they are not independently associated ([Figure 3.23](#)). There were no secondary associations in the *ABO* locus when conditioned on the lead *ABO* SNP ([Figure 3.23](#)).

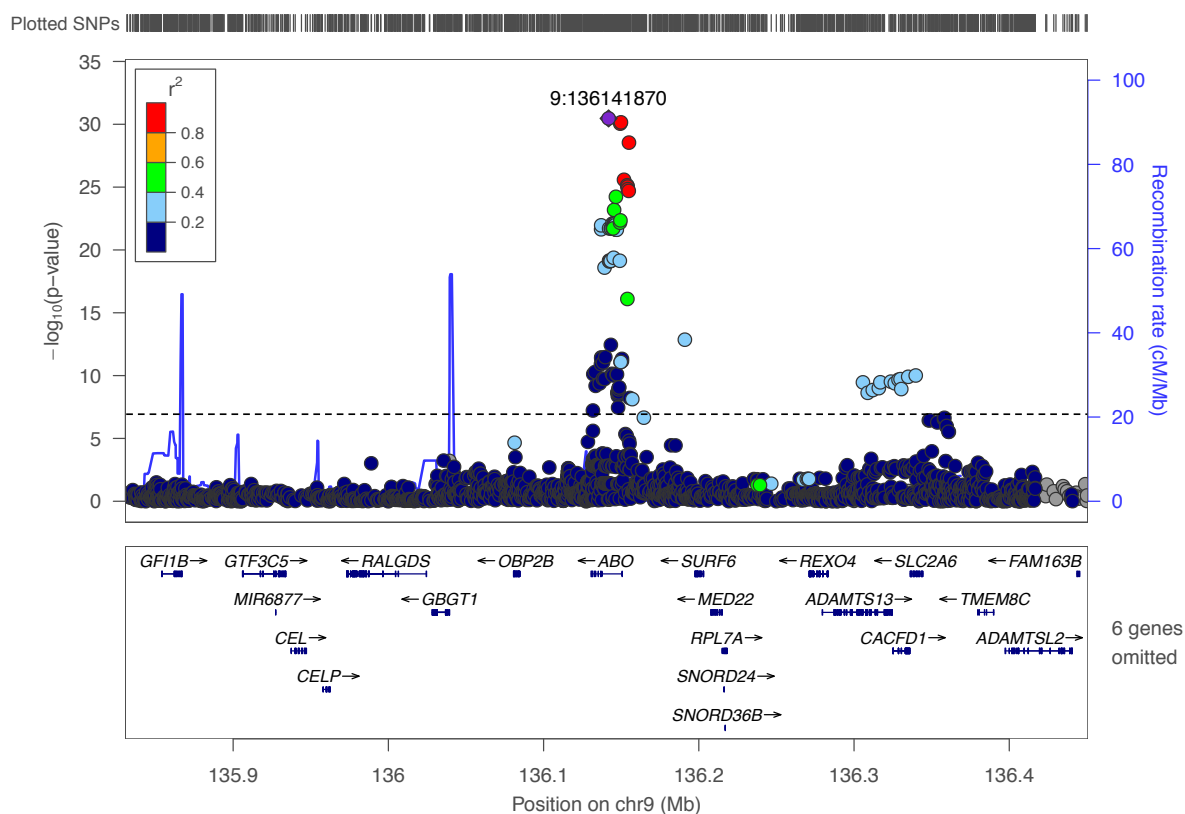


Figure 3.21 Regional association plot of the associated locus in chromosome 9

SNP p -values are plotted against genomic position and were generated with the association testing described in [Figure 3.19](#). The most statistically significant SNP (9:136141870=rs2519093) is plotted in purple with the correlation (linkage disequilibrium, from 1000 Genomes and HapMap) with respect to it, shown on a colour scale from red (high) to dark blue (low). Gene positions are shown on the bottom panel of each plot. The recombination rates (from HapMap) are plotted on the right axis. The regional association plot was generated using Locuszoom.(239) Some gene track annotations have been omitted to improve visualisation.

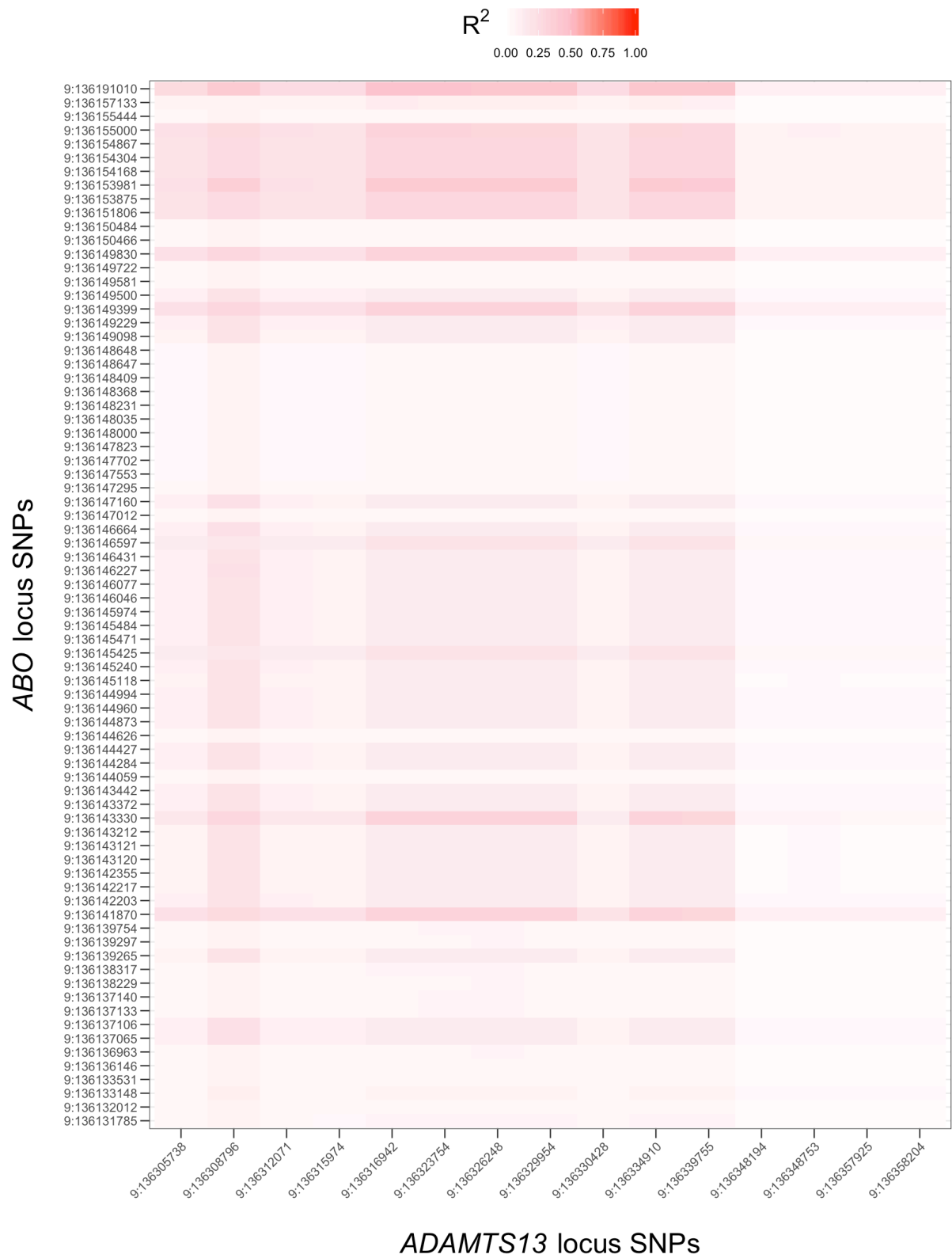
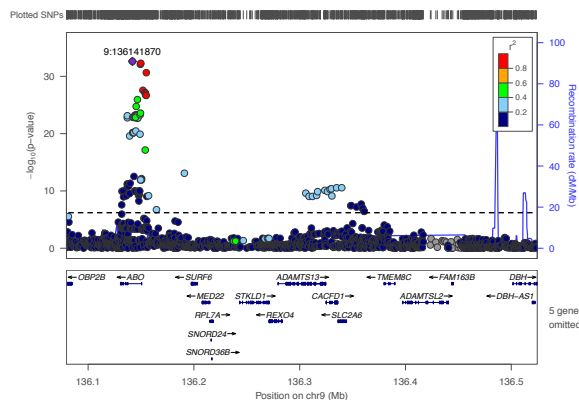


Figure 3.22 Linkage disequilibrium heat maps of significant SNPs in the *ABO* and *ADAMTS13* loci

Pairwise SNP correlations for the significant SNPs in chromosome 9 were used to calculate LD using PLINK for the parameters R^2 . SNPs were assigned to the *ABO* or the *ADAMTS13* locus depending on their proximity to the nearest gene. Each axis tick mark represents a SNP (displayed as CHR:POS). The LD range is from low LD (0; no correlation) to high LD (1; perfect correlation).

A Associated loci



B Conditional analysis

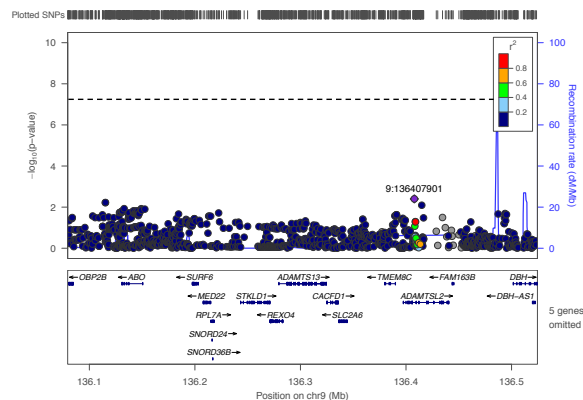


Figure 3.23 Conditional analysis at the associated chromosome 9 locus

A Regional association plot of the significant locus from the analysis described in [Figure 3.19](#) and shown in [Figure 3.21](#).

B The analysis was repeated and conditioned on the lead SNP in *ABO* to identify independent associations (case/control group ~ SNP allelic dosage + lead *ABO* SNP (rs2519093) + EV1+ ... + EV5). The dotted horizontal lines represent a p -value threshold of 5×10^{-8} .

3.2.3.2 Discovery cohort analysis

The discovery cohort included 678 CTEPH cases from UK centres (Papworth and Imperial) and 1492 healthy controls following quality control and imputation. The data was divided into discovery and validation cohorts from the combined datasets using a *priori* group definitions. Association testing was performed as described for the joint analysis using logistic regression with post-imputation SNP dosages ($n=7,675,738$) assuming an additive model and adjusted for 5 principal components. There were 68

associated SNPs in chromosomes 4 and 9 (Figure 3.24 and Table 3.7). The lead SNP in chromosome 4 (rs2036914, OR (95% CI) = 1.43 (1.30-1.56), $p=4.79 \times 10^{-8}$) was an intronic variant in the *F11* gene. The associated locus in chromosome 9 was the same as for the joint analysis, with the same lead intronic SNP (rs2519093) in *ABO*.

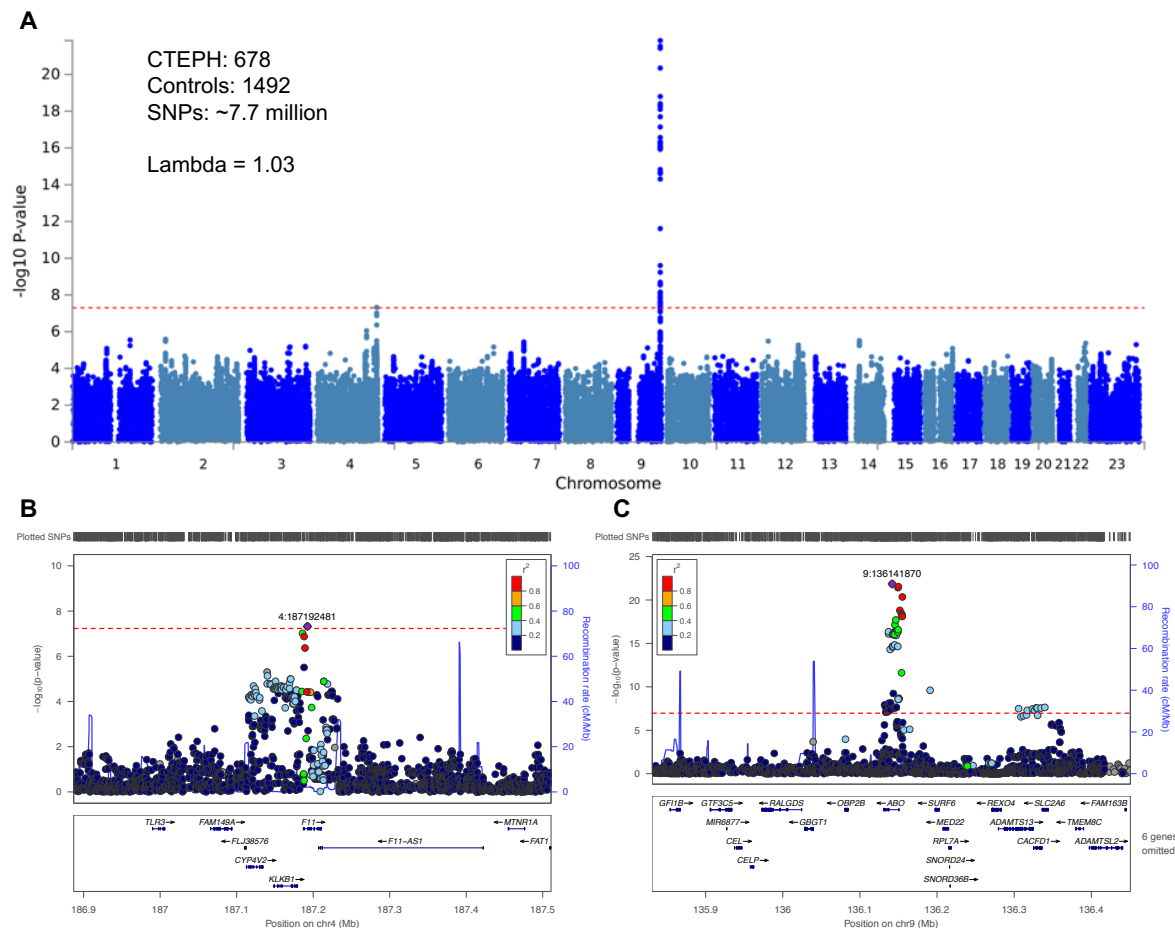


Figure 3.24 Associated loci in the discovery cohort

Association testing for 678 CTEPH cases and 1492 healthy controls. Logistic regression was performed using post-imputation SNP dosages ($n=7,675,738$) assuming an additive model and adjusted for 5 ancestry informative principal components.

A Manhattan plot of the discovery cohort associations

B Regional association plot of the chromosome 4 locus in proximity to the *F11* gene

C Regional association plot of the chromosome 9 locus in proximity to the *ABO* gene

Some gene tracks are omitted to improve visualisation

rsID	CHR:POS_EA/NEA	GENE	FUNC	EAF_A	EAF_U	EAF_REF	INFO	OR (95% CI)	p
rs2519093	9:136141870_T/C	ABO	intronic	0.298	0.165	0.182	0.951	2.22 (2.06-2.38)	1.42e-22
rs532436	9:136149830_A/G	ABO	intronic	0.297	0.165	0.185	0.965	2.21 (2.05-2.37)	2.89e-22
rs507666	9:136149399_A/G	ABO	intronic	0.297	0.166	0.185	0.965	2.2 (2.04-2.36)	3.77e-22
rs635634*	9:136155000_T/C	ABO	intergenic	0.302	0.174	0.185	0.977	2.14 (1.98-2.29)	4.47e-21
rs600038*	9:136151806_C/T	ABO	intergenic	0.316	0.189	0.218	0.965	2.04 (1.88-2.19)	1.57e-19
rs651007*	9:136153875_T/C	ABO	intergenic	0.316	0.189	0.215	0.971	2 (1.85-2.15)	3.86e-19
rs579459*	9:136154168_C/T	ABO	intergenic	0.316	0.189	0.215	1.000	1.99 (1.84-2.14)	4.47e-19
rs649129	9:136154304_T/C	ABO	intergenic	0.316	0.189	0.215	0.988	2 (1.85-2.15)	5.61e-19
rs495828	9:136154867_T/G	ABO	intergenic	0.316	0.189	0.215	0.964	2 (1.85-2.16)	7.95e-19
rs550057	9:136146597_T/C	ABO	intronic	0.370	0.241	0.280	0.972	1.93 (1.79-2.08)	2e-18
rs9411378	9:136145425_A/C	ABO	intronic	0.364	0.236	0.290	0.887	1.99 (1.83-2.14)	7.19e-18
rs529565	9:136149500_C/T	ABO	intronic	0.448	0.313	0.345	0.976	1.82 (1.68-1.96)	2.67e-17
rs687289*	9:136137106_A/G	ABO	intronic	0.448	0.315	0.353	0.989	1.81 (1.67-1.95)	4.53e-17
rs2769071	9:136145974_G/A	ABO	intronic	0.448	0.315	0.350	0.971	1.82 (1.68-1.96)	4.77e-17
rs505922*	9:136149229_C/T	ABO	intronic	0.447	0.313	0.345	1.000	1.8 (1.66-1.93)	4.82e-17
rs677355	9:136146046_A/G	ABO	intronic	0.448	0.315	0.353	0.971	1.82 (1.68-1.96)	5e-17
rs597988	9:136144284_A/T	ABO	intronic	0.448	0.315	0.348	0.992	1.8 (1.67-1.94)	5.1e-17
rs492488	9:136144960_A/G	ABO	intronic	0.448	0.315	0.350	0.989	1.81 (1.67-1.94)	5.15e-17
rs582118	9:136145471_G/A	ABO	intronic	0.448	0.315	0.350	0.991	1.8 (1.67-1.94)	5.21e-17
rs582094	9:136145484_T/A	ABO	intronic	0.448	0.315	0.350	0.990	1.8 (1.67-1.94)	5.27e-17
rs676996	9:136146077_G/T	ABO	intronic	0.448	0.315	0.353	0.991	1.8 (1.67-1.94)	5.38e-17
rs676457	9:136146227_T/A	ABO	intronic	0.447	0.315	0.350	0.991	1.8 (1.66-1.94)	6.59e-17

rs687621*	9:136137065_G/A	ABO	intronic	0.448	0.315	0.350	1.000	1.79 (1.65-1.93)	7.91e-17
rs493246	9:136144994_A/G	ABO	intronic	0.448	0.315	0.350	0.999	1.79 (1.65-1.93)	9.06e-17
rs495203	9:136145240_T/C	ABO	intronic	0.448	0.315	0.350	0.999	1.79 (1.65-1.93)	9.14e-17
rs514659*	9:136142203_C/A	ABO	intronic	0.448	0.315	0.350	1.000	1.79 (1.65-1.92)	9.26e-17
rs545971	9:136143372_T/C	ABO	intronic	0.448	0.315	0.350	1.000	1.79 (1.65-1.92)	9.32e-17
rs612169	9:136143442_G/A	ABO	intronic	0.448	0.315	0.348	1.000	1.79 (1.65-1.92)	9.32e-17
rs8176663	9:136144427_C/T	ABO	intronic	0.448	0.315	0.350	1.000	1.79 (1.65-1.92)	9.32e-17
rs491626	9:136144873_T/C	ABO	intronic	0.448	0.315	0.350	1.000	1.79 (1.65-1.92)	9.32e-17
rs527210	9:136146431_T/C	ABO	intronic	0.446	0.315	0.350	0.977	1.8 (1.66-1.94)	1.06e-16
rs674302	9:136146664_A/T	ABO	intronic	0.447	0.315	0.350	0.999	1.79 (1.65-1.92)	1.17e-16
rs554833	9:136147160_T/C	ABO	intronic	0.447	0.315	0.350	0.997	1.79 (1.65-1.92)	1.17e-16
rs494242	9:136145118_T/C	ABO	intronic	0.466	0.339	0.377	0.985	1.74 (1.6-1.88)	1.48e-15
rs644234	9:136142217_G/T	ABO	intronic	0.466	0.339	0.377	0.994	1.73 (1.59-1.86)	2.11e-15
rs8176645	9:136149098_A/T	ABO	intronic	0.451	0.322	0.375	0.735	1.95 (1.78-2.11)	2.23e-15
rs613534	9:136143120_G/A	ABO	intronic	0.466	0.339	0.377	0.995	1.73 (1.59-1.86)	2.29e-15
rs543968	9:136143121_C/T	ABO	intronic	0.466	0.339	0.377	0.995	1.73 (1.59-1.86)	2.29e-15
rs544873	9:136143212_A/G	ABO	intronic	0.466	0.339	0.377	0.995	1.73 (1.59-1.86)	2.29e-15
rs643434	9:136142355_A/G	ABO	intronic	0.466	0.339	0.377	0.996	1.72 (1.59-1.86)	2.47e-15
rs657152*	9:136139265_A/C	ABO	intronic	0.466	0.339	0.377	1.000	1.71 (1.58-1.85)	4.92e-15
rs11244061	9:136153981_T/C	ABO	intergenic	0.176	0.102	0.120	0.947	1.97 (1.78-2.16)	2.44e-12
rs11244084	9:136191010_T/C	LCN1P2	intergenic	0.159	0.091	0.075	0.850	1.92 (1.72-2.13)	2.5e-10
rs142956930	9:136143330_G/A	ABO	intronic	0.125	0.066	0.017	0.513	2.99 (2.65-3.34)	5.83e-10
rs559723	9:136150484_A/G	ABO	intronic	0.392	0.490	0.498	0.988	1.51 (1.37-1.64)	2.03e-09

rs616154	9:136150466_C/T	<i>ABO</i>	intronic	0.392	0.491	0.500	0.976	1.51 (1.37-1.64)	2.64e-09
rs630014*	9:136149722_A/G	<i>ABO</i>	intronic	0.391	0.489	0.495	1.000	1.5 (1.36-1.63)	2.69e-09
rs630510	9:136149581_A/G	<i>ABO</i>	intronic	0.391	0.489	0.495	0.998	1.5 (1.36-1.63)	2.86e-09
rs8176681	9:136139754_T/C	<i>ABO</i>	intronic	0.642	0.548	0.592	0.990	1.49 (1.35-1.62)	6.94e-09
rs2073827	9:136137133_G/C	<i>ABO</i>	intronic	0.642	0.548	0.592	0.989	1.49 (1.35-1.62)	7.75e-09
rs2073828	9:136137140_G/A	<i>ABO</i>	intronic	0.642	0.548	0.595	0.989	1.49 (1.35-1.62)	8.7e-09
rs7873635	9:136132012_T/C	<i>ABO</i>	intronic	0.340	0.428	0.383	0.831	1.57 (1.41-1.72)	1.18e-08
rs8176690	9:136138317_A/G	<i>ABO</i>	intronic	0.642	0.548	0.592	0.980	1.48 (1.34-1.61)	1.34e-08
rs2073826	9:136136963_G/T	<i>ABO</i>	intronic	0.642	0.548	0.595	0.977	1.48 (1.34-1.61)	1.59e-08
rs8176715	9:136133148_T/C	<i>ABO</i>	intronic	0.387	0.481	0.410	0.917	1.49 (1.35-1.63)	2.25e-08
rs8176649	9:136147295_G/A	<i>ABO</i>	intronic	0.661	0.571	0.620	0.992	1.47 (1.34-1.61)	2.31e-08
rs3124761	9:136339755_C/T	<i>SLC2A6</i>	intronic	0.808	0.877	0.838	0.876	1.7 (1.52-1.89)	2.33e-08
rs7046674	9:136147012_C/T	<i>ABO</i>	intronic	0.661	0.571	0.620	0.991	1.47 (1.34-1.61)	2.34e-08
rs8176668	9:136144059_A/T	<i>ABO</i>	intronic	0.661	0.572	0.617	0.977	1.47 (1.34-1.61)	2.41e-08
rs3124765	9:136328657_C/T	<i>CACFD1</i>	exonic	0.808	0.877	0.839	0.892	1.7 (1.51-1.88)	2.59e-08
rs3124764	9:136329954_C/T	<i>CACFD1</i>	intronic	0.808	0.877	0.840	0.892	1.7 (1.51-1.88)	2.75e-08
rs3094379	9:136334910_C/T	<i>CACFD1</i>	UTR3	0.808	0.876	0.838	0.887	1.69 (1.51-1.88)	2.75e-08
rs7036642	9:136144626_G/A	<i>ABO</i>	intronic	0.661	0.572	0.620	0.985	1.47 (1.33-1.61)	2.78e-08
rs28645493*	9:136305738_G/C	<i>ADAMTS13</i>	intronic	0.131	0.076	0.098	1.000	1.81 (1.6-2.02)	3.27e-08
rs4962153*	9:136323754_G/A	<i>ADAMTS13</i>	intronic	0.808	0.877	0.838	0.900	1.68 (1.5-1.87)	3.54e-08
rs8176682*	9:136139297_C/T	<i>ABO</i>	intronic	0.660	0.572	0.620	0.990	1.46 (1.32-1.6)	4.52e-08
rs8176691*	9:136138229_C/T	<i>ABO</i>	intronic	0.660	0.572	0.620	0.989	1.46 (1.32-1.6)	4.77e-08
rs2036914	4:187192481_T/C	<i>F11</i>	intronic	0.386	0.479	0.485	1.000	1.43 (1.3-1.56)	4.79e-08

Table 3.7 Significant SNPs in the discovery cohort

Association testing for 678 CTEPH cases and 1492 healthy controls. Logistic regression was performed using post-imputation SNP dosages (n=7,675,738) assuming an additive model and adjusted for 5 ancestry informative principal components. * SNPs with an asterisk were present on the micro-array chip pre-imputation and those without have been imputed. The column headings and additional details are described in [Table 3.6](#).

3.2.3.3 Validation cohort analysis

The validation cohort included 572 CTEPH cases from other European and US centres and the same 1492 healthy controls following quality control and imputation. Association testing was performed as described for the joint analysis and discovery cohort. There were 37 associated SNPs in chromosome 9 ([Figure 3.25](#) and [Table 3.8](#)). The associated locus in chromosome 9 was the same as for the joint analysis and discovery cohort, with the same lead intronic SNP (rs2519093) in *ABO*. The odds ratio for the lead SNP is higher in the validation cohort than in the discovery cohort (OR: 2.2 vs. 2.7) and there is a corresponding higher effect allele frequency in the CTEPH group (EAF: 0.337 vs. 0.298). The significant chromosome 4 locus identified in the discovery cohort was not replicated in the validation cohort.

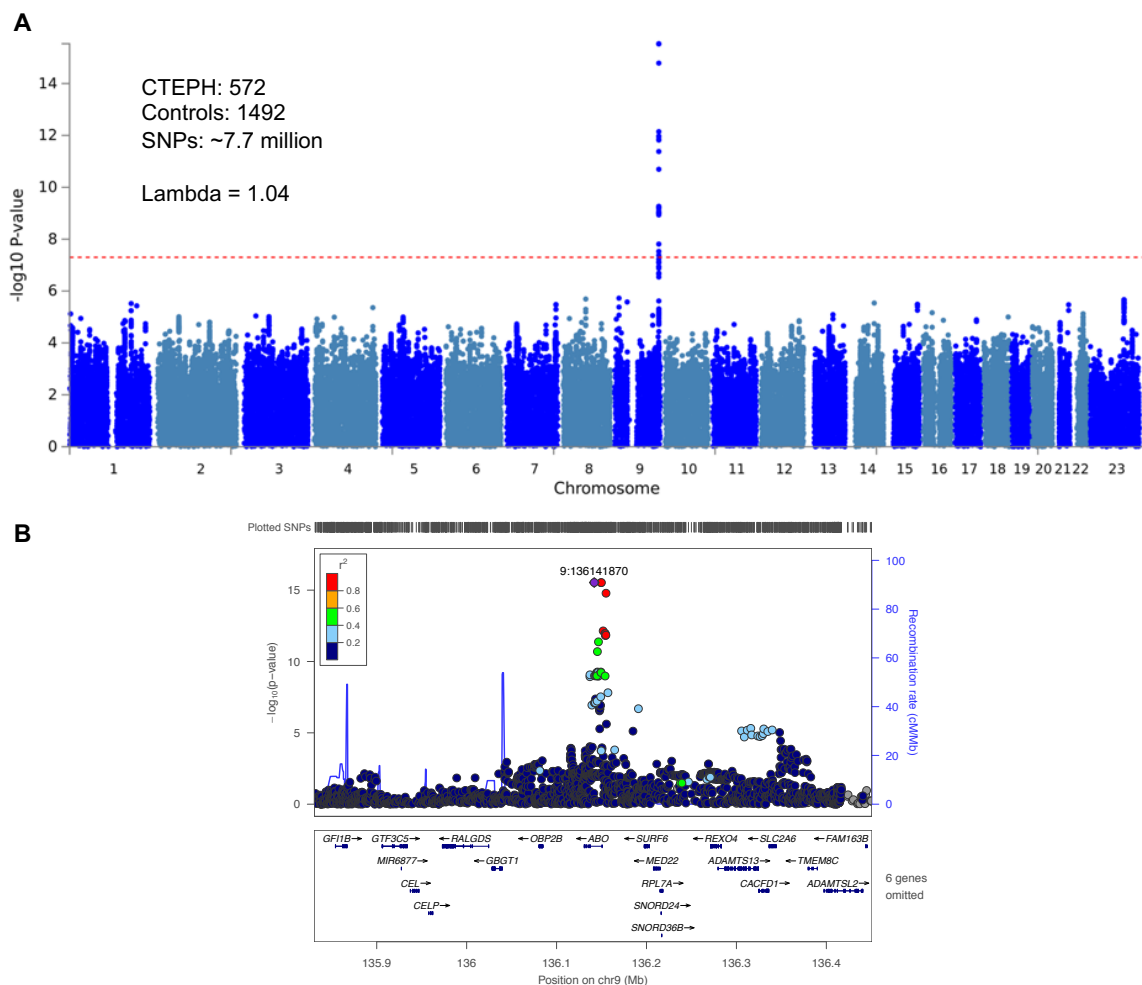


Figure 3.25 Associated loci in the validation cohort

Association testing for 572 CTEPH cases and 1492 healthy controls. Logistic regression was performed using post-imputation SNP dosages (n=7,675,738) assuming an additive model and adjusted for 5 ancestry informative principal components.

A Manhattan plot of the discovery cohort associations

B Regional association plot of the chromosome 9 locus in proximity to the ABO gene

Some gene tracks are omitted to improve visualisation.

rsID	CHR:POS_EA/NEA	GENE	FUNC	EAFA	EAU	EA_REF	INFO	OR (95% CI)	p
rs2519093	9:136141870_T/C	ABO	intronic	0.337	0.165	0.182	0.951	2.72 (2.48-2.96)	2.92e-16
rs532436	9:136149830_A/G	ABO	intronic	0.338	0.165	0.185	0.965	2.72 (2.48-2.96)	2.95e-16
rs507666	9:136149399_A/G	ABO	intronic	0.338	0.166	0.185	0.965	2.72 (2.48-2.96)	2.98e-16
rs635634*	9:136155000_T/C	ABO	intergenic	0.343	0.174	0.185	0.977	2.65 (2.41-2.89)	1.64e-15
rs600038	9:136151806_C/T	ABO	intergenic	0.348	0.189	0.218	0.965	2.37 (2.14-2.61)	7.18e-13
rs651007*	9:136153875_T/C	ABO	intergenic	0.348	0.189	0.215	0.971	2.34 (2.1-2.57)	1.07e-12
rs579459*	9:136154168_C/T	ABO	intergenic	0.348	0.189	0.215	1.000	2.32 (2.09-2.55)	1.11e-12
rs495828	9:136154867_T/G	ABO	intergenic	0.348	0.189	0.215	0.964	2.34 (2.11-2.58)	1.43e-12
rs649129	9:136154304_T/C	ABO	intergenic	0.347	0.189	0.215	0.988	2.33 (2.09-2.56)	1.51e-12
rs550057	9:136146597_T/C	ABO	intronic	0.405	0.241	0.280	0.972	2.22 (2-2.45)	4.19e-12
rs9411378	9:136145425_A/C	ABO	intronic	0.398	0.236	0.290	0.887	2.28 (2.04-2.52)	2.02e-11
rs582094	9:136145484_T/A	ABO	intronic	0.482	0.315	0.350	0.990	1.95 (1.74-2.16)	5.44e-10
rs505922*	9:136149229_C/T	ABO	intronic	0.482	0.313	0.345	1.000	1.94 (1.73-2.15)	5.45e-10
rs2769071	9:136145974_G/A	ABO	intronic	0.482	0.315	0.350	0.971	1.96 (1.75-2.18)	5.82e-10
rs677355	9:136146046_A/G	ABO	intronic	0.482	0.315	0.353	0.971	1.96 (1.75-2.18)	5.83e-10
rs676996	9:136146077_G/T	ABO	intronic	0.482	0.315	0.353	0.991	1.95 (1.73-2.16)	5.84e-10
rs676457	9:136146227_T/A	ABO	intronic	0.482	0.315	0.350	0.991	1.95 (1.73-2.16)	5.84e-10
rs582118	9:136145471_G/A	ABO	intronic	0.482	0.315	0.350	0.991	1.94 (1.73-2.16)	5.88e-10
rs529565	9:136149500_C/T	ABO	intronic	0.481	0.313	0.345	0.976	1.95 (1.74-2.16)	5.99e-10
rs597988	9:136144284_A/T	ABO	intronic	0.482	0.315	0.348	0.992	1.94 (1.73-2.15)	6.37e-10
rs492488	9:136144960_A/G	ABO	intronic	0.482	0.315	0.350	0.989	1.94 (1.73-2.15)	6.81e-10

rs687289*	9:136137106_A/G	ABO	intronic	0.482	0.315	0.353	0.989	1.94 (1.73-2.15)	8.56e-10
rs674302	9:136146664_A/T	ABO	intronic	0.482	0.315	0.350	0.999	1.92 (1.71-2.13)	9.11e-10
rs493246	9:136144994_A/G	ABO	intronic	0.482	0.315	0.350	0.999	1.92 (1.71-2.13)	9.16e-10
rs495203	9:136145240_T/C	ABO	intronic	0.482	0.315	0.350	0.999	1.92 (1.71-2.13)	9.17e-10
rs554833	9:136147160_T/C	ABO	intronic	0.482	0.315	0.350	0.997	1.92 (1.71-2.13)	9.53e-10
rs514659*	9:136142203_C/A	ABO	intronic	0.482	0.315	0.350	1.000	1.92 (1.71-2.13)	9.97e-10
rs545971	9:136143372_T/C	ABO	intronic	0.482	0.315	0.350	1.000	1.92 (1.71-2.13)	1e-09
rs612169	9:136143442_G/A	ABO	intronic	0.482	0.315	0.348	1.000	1.92 (1.71-2.13)	1e-09
rs8176663	9:136144427_C/T	ABO	intronic	0.482	0.315	0.350	1.000	1.92 (1.71-2.13)	1e-09
rs491626	9:136144873_T/C	ABO	intronic	0.482	0.315	0.350	1.000	1.92 (1.71-2.13)	1e-09
rs11244061	9:136153981_T/C	ABO	intergenic	0.199	0.102	0.120	0.947	2.43 (2.15-2.72)	1.04e-09
rs527210	9:136146431_T/C	ABO	intronic	0.481	0.315	0.350	0.977	1.93 (1.72-2.15)	1.08e-09
rs687621*	9:136137065_G/A	ABO	intronic	0.482	0.315	0.350	1.000	1.92 (1.71-2.13)	1.16e-09
rs558240*	9:136157133_A/G	ABO	intergenic	0.511	0.402	0.405	1.000	1.77 (1.58-1.97)	1.54e-08
rs8176645	9:136149098_A/T	ABO	intronic	0.478	0.322	0.375	0.735	2.05 (1.8-2.3)	2.98e-08
rs142956930	9:136143330_G/A	ABO	intronic	0.139	0.066	0.017	0.513	4.35 (3.82-4.87)	4.18e-08

Table 3.8 Significant SNPs in the validation cohort

Association testing for 572 CTEPH cases and 1492 healthy controls. Logistic regression was performed using post-imputation SNP dosages (n=7,675,738) assuming an additive model and adjusted for 5 ancestry informative principal components. * SNPs with an asterisk were present on the micro-array chip pre-imputation and those without have been imputed. The column headings and additional details are described in [Table 3.6](#).

3.2.3.4 Genotyping quality of the significant GWAS associations

To confirm that the significant associations described in the *ABO* locus and the *F11* locus (for the discovery cohort analysis) were adequately genotyped, the micro-array clustering plots were examined. Prior to imputation, association testing was performed in a joint analysis of all samples (CTEPH cases=1250, Healthy controls=1492) and 640,744 SNPs (275,255 SNPs were excluded by a MAF threshold of <1%). Logistic regression adjusted for 5 ancestry informative principal components identified the lead SNP in these two regions as rs635634 (9:136155000_T/C, OR 2.34, $p=5.71 \times 10^{-30}$) and rs2289252 (4:187207381_T/C, OR=1.34, $p=7.97 \times 10^{-7}$) ([Table 3.9](#) and [Figure 3.26](#)). Micro-array clustering was adequate for these two SNPs, which confirmed that the associations were not due to genotyping errors ([Figure 3.27](#)).

rsID	CHR:POS_EA/NEA	GENE	FUNC	OR (95% CI)	P
rs635634	9:136155000_T/C	<i>ABO</i>	intergenic	2.34 (2.19-2.48)	5.71e-30
rs651007	9:136153875_T/C	<i>ABO</i>	intergenic	2.12 (1.98-2.25)	9.08e-26
rs579459	9:136154168_C/T	<i>ABO</i>	intergenic	2.12 (1.98-2.25)	9.08e-26
rs505922	9:136149229_C/T	<i>ABO</i>	intronic	1.88 (1.75-2)	6.75e-23
rs687289	9:136137106_A/G	<i>ABO</i>	intronic	1.86 (1.73-1.98)	3.31e-22
rs514659	9:136142203_C/A	<i>ABO</i>	intronic	1.86 (1.74-1.99)	1.92e-22
rs687621	9:136137065_G/A	<i>ABO</i>	intronic	1.86 (1.74-1.99)	2.27e-22
rs657152	9:136139265_A/C	<i>ABO</i>	intronic	1.76 (1.63-1.88)	2.54e-19
rs630014	9:136149722_A/G	<i>ABO</i>	intronic	1.52 (1.4-1.64)	8.23e-12
rs8176682	9:136139297_C/T	<i>ABO</i>	intronic	1.48 (1.36-1.6)	2.18e-10
rs4962153	9:136323754_G/A	<i>ADAMTS13</i>	intronic	1.66 (1.5-1.82)	5.44e-10
rs28645493	9:136305738_G/C	<i>ADAMTS13</i>	intronic	1.86 (1.66-2.05)	3.44e-10
rs558240	9:136157133_A/G	<i>ABO</i>	intergenic	1.4 (1.29-1.52)	8.74e-09

Table 3.9 Significant SNPs in the pre-imputation GWAS analysis

Association testing for 1250 CTEPH cases and 1492 healthy controls prior to imputation using the SNPs available on the original micro-array chip (n=640,744 SNPs following QC). Logistic regression was performed using pre-imputation SNPs assuming an additive model and adjusted for 5 ancestry informative principal components. The column headings and additional details are described in [Table 3.6](#).

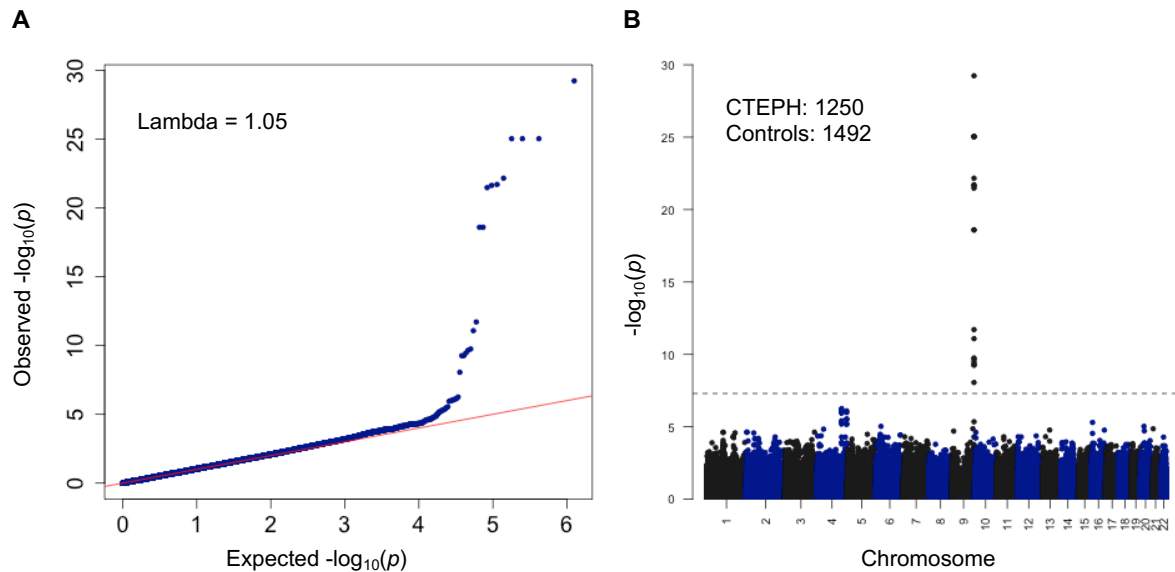


Figure 3.26 Case-control association testing pre-imputation

Analysis of 1250 CTEPH cases (discovery and validation cohorts combined), 1492 healthy controls and 640,744 SNPs. Logistic regression was performed using pre-imputation SNPs assuming an additive model and adjusted for 5 ancestry informative principal components. A p -value of $<5 \times 10^{-8}$ was considered genome-wide significant (dotted grey line **B**). Genomic inflation factor (λ)=1.05.

A Quantile-quantile (QQ) plot of the observed and expected p -values.

B Manhattan plot of p -values plotted against genomic position.

P -values are transformed to a $-\log_{10}$ scale.

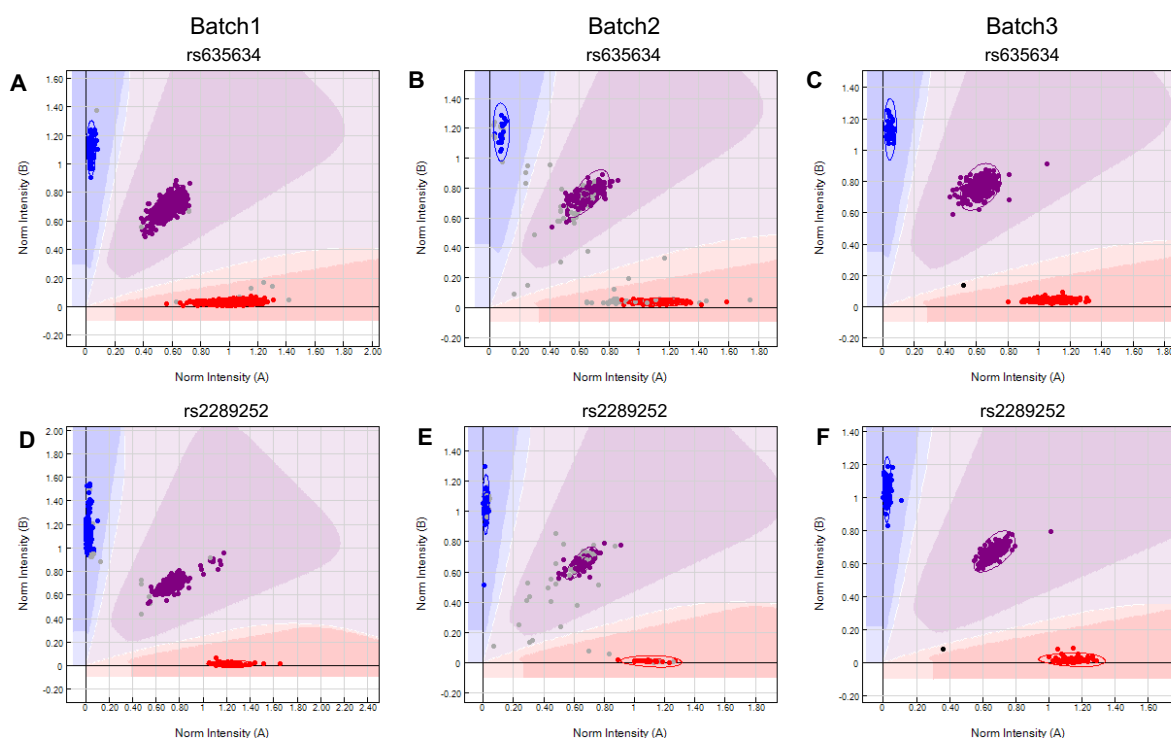


Figure 3.27 Micro-array clustering plots for the lead SNP associations

A, B and C micro-array cluster plots for the lead SNP association rs635634 in chromosome 9 prior to imputation from a joint analysis.

D, E and F micro-array cluster plots for the lead SNP association rs2289252 in chromosome 9 prior to imputation from a joint analysis.

Micro-array plot interpretation is described in [Figure 3.9](#).

3.2.4 The *ABO* association

The *ABO* locus was associated with CTEPH in the discovery, validation and joint analysis. *ABO* groups have been associated with CTEPH in international observational studies with an over-representation of non-O blood groups.(68) There are different risk associations within the non-O blood groups for venous thromboembolism. The A1 subtype has a higher VTE risk association than the A2 subgroup, however both would be classified as blood group A on serological testing.(253) Reconstructing genetic *ABO* groups enabled investigation of CTEPH associations with comprehensive *ABO* subgroups.

The allele frequencies of the 4 “tagging” SNPs used to reconstruct the ABO subgroups are shown in [Table 3.10](#). The SNP associated with the A1 ABO group (rs507666) is over-represented in CTEPH compared with study healthy controls and a European (non-Finnish) reference population (EAF: 0.316 vs. 0.166 vs. 0.185 respectively) ([Table 3.10](#)). This SNP is also the third most significant association in the joint analysis (OR (95% CI): 2.38 (2.23-2.53), $p=8.89 \times 10^{-31}$) and highly correlated with the lead SNP (LDlink: $R^2 = 0.991$, $p<0.001$, European (non-Finnish)). Conversely, the SNP that tags the O ABO group is under-represented in CTEPH compared with healthy controls and a reference population (EAF: 0.537 vs. 0.685 vs. 0.647 respectively) ([Table 3.10](#)). The SNPs tagging the A2 and B subgroups have similar allele frequencies in CTEPH and healthy controls.

Reconstructing ABO subgroups resulted in 10 genotype groups (A1A1, A1A2, A1B, A1O, A2A2, A2B, A2O, BB, BO, OO), from which blood groups A, B, AB and O were inferred. The A1 group was enriched in CTEPH compared with healthy controls ([Figure 3.28A](#)). The inferred A blood group occurred in 719 (59%) CTEPH cases and 567 (38%) healthy controls ([Figure 3.28B](#)). The inferred O blood group was under-represented in CTEPH (317 (26%)) compared with healthy controls (697 (47%)).

The risk of CTEPH differed within the comprehensive genetic non-O groups with the highest risk in the A1A1 group (OR (95% CI) 4.39 (2.92-6.69), $p<0.001$) and the subgroups enriched by A1 ([Figure 3.29A](#)). Interestingly, the risk of CTEPH was not increased in the largest A2 enriched group (A2O: OR (95% CI) 1.11 (0.80-1.53), $p=0.544$) but was increased in the equivalent A1 group (A1O: 3.04 (2.46-3.75), $p<0.001$). The risk of CTEPH was increased in the largest B enriched subgroup (BO: 1.66 (1.22-2.27), $p<0.001$) but the magnitude was less than the A1 enriched groups. The risk of CTEPH was increased in the A, B and AB groups when they were inferred from the 10 comprehensive genetic ABO genotypic groups ([Figure 3.29B](#)). The differential effects of A1 and A2 on CTEPH risk results in a lower odds ratio for the A groups (containing A1 and A2) than described for A1 enriched subgroups in [Figure 3.29A](#).

Genetic <i>ABO</i> group	rsID	CHR:POS_EA/NEA	EAf_A	EAf_U	EAf_REF	OR (95% CI)	<i>p</i>
O	rs687289	9:136137106_G/A	0.537	0.685	0.647	0.53 (0.50-0.57)	1.10e-22
A1	rs507666	9:136149399_A/G	0.316	0.166	0.185	2.38 (2.23-2.53)	8.89e-31
A2	rs8176704	9:136135552_A/G	0.062	0.072	0.090	0.88 (0.64-1.13)	0.324
B	rs8176746	9:136131322_T/G	0.077	0.073	0.066	1.04 (0.82-1.27)	0.711

Table 3.10 Effect allele frequencies for the tagging SNPs used to reconstruct *ABO* subgroups

The odds ratios and allele frequencies are for the effect allele which tags the *ABO* subgroup (see [Table 2.1](#) Material and Methods). Consequently, for rs687289 the effect and non-effect alleles are reversed compared to those reported in [Tables 3.6 - 3.8](#). The column headings and additional details are described in [Table 3.6](#).

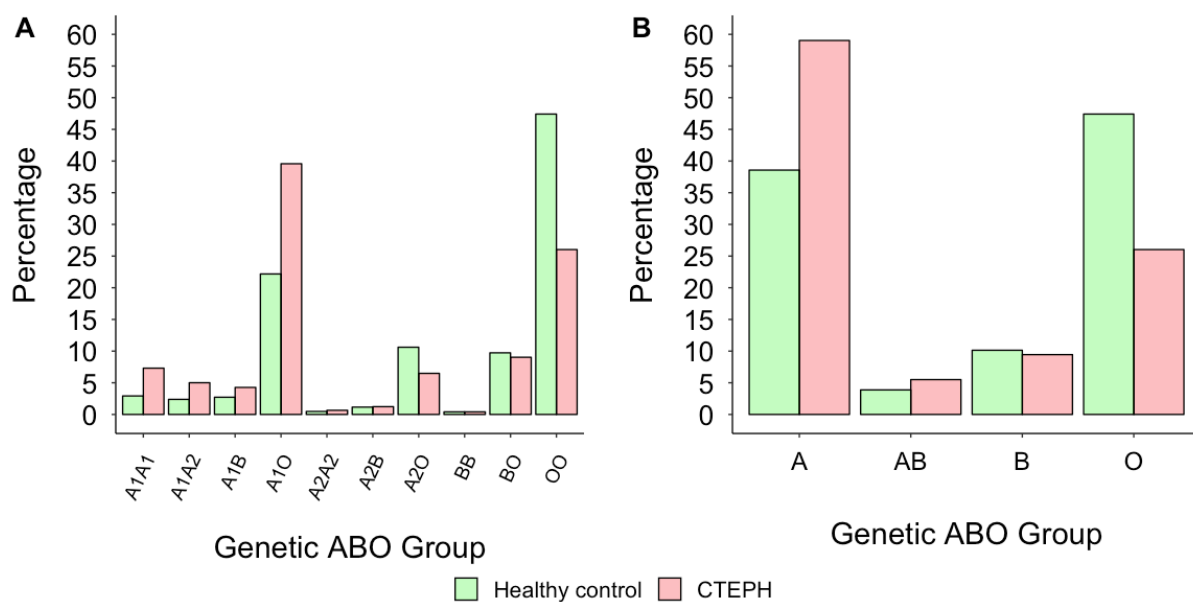


Figure 3.28 Genetic ABO groups in CTEPH and healthy controls

A The percentages of the comprehensive genotypic ABO groups

B The percentages of the inferred A, AB, B and O ABO groups

The numbers in each group are shown in [Figure 3.29](#).

ABO group frequencies can vary by populations and subpopulations.(254) To investigate whether the ABO association was being driven by allele frequency differences in certain subpopulations, the genetic ABO group frequencies were subdivided by centre ([Figure 3.30](#)). In CTEPH, the over representation of the A group and the under representation of the O ABO group was consistent across all centres.

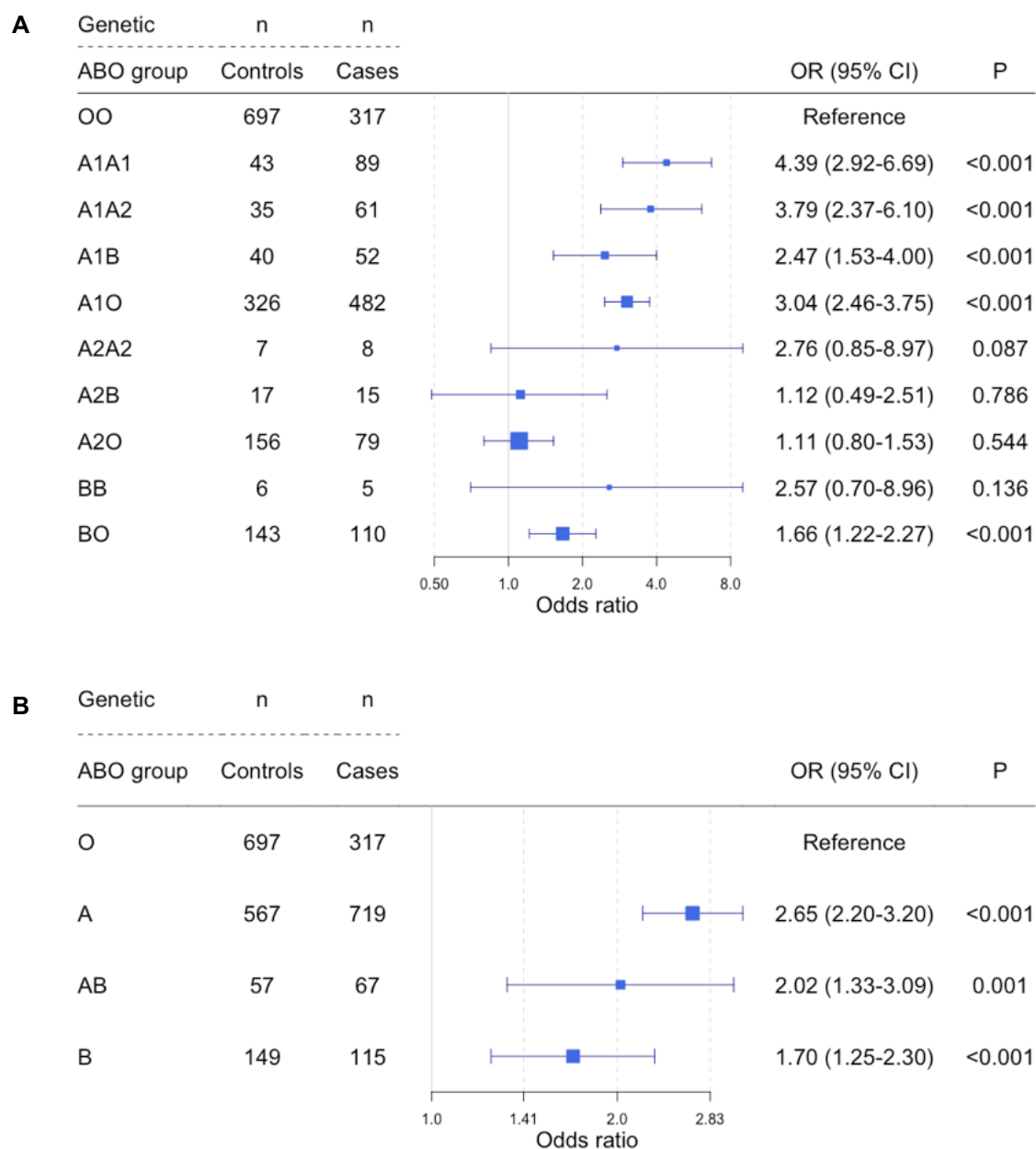


Figure 3.29 Genetic ABO groups and CTEPH risk

Odds ratios for CTEPH (with respect to healthy controls) in different genetic ABO groups were calculated using logistic regression and adjusted for 5 ancestry informative principal components.

A Comprehensive genetic ABO groups

B 4 inferred genetic ABO groups (O, A, AB and B)

Of 1492 healthy controls and 1250 CTEPH cases, genetic ABO groups could not be inferred for n=22 and n=32 respectively.

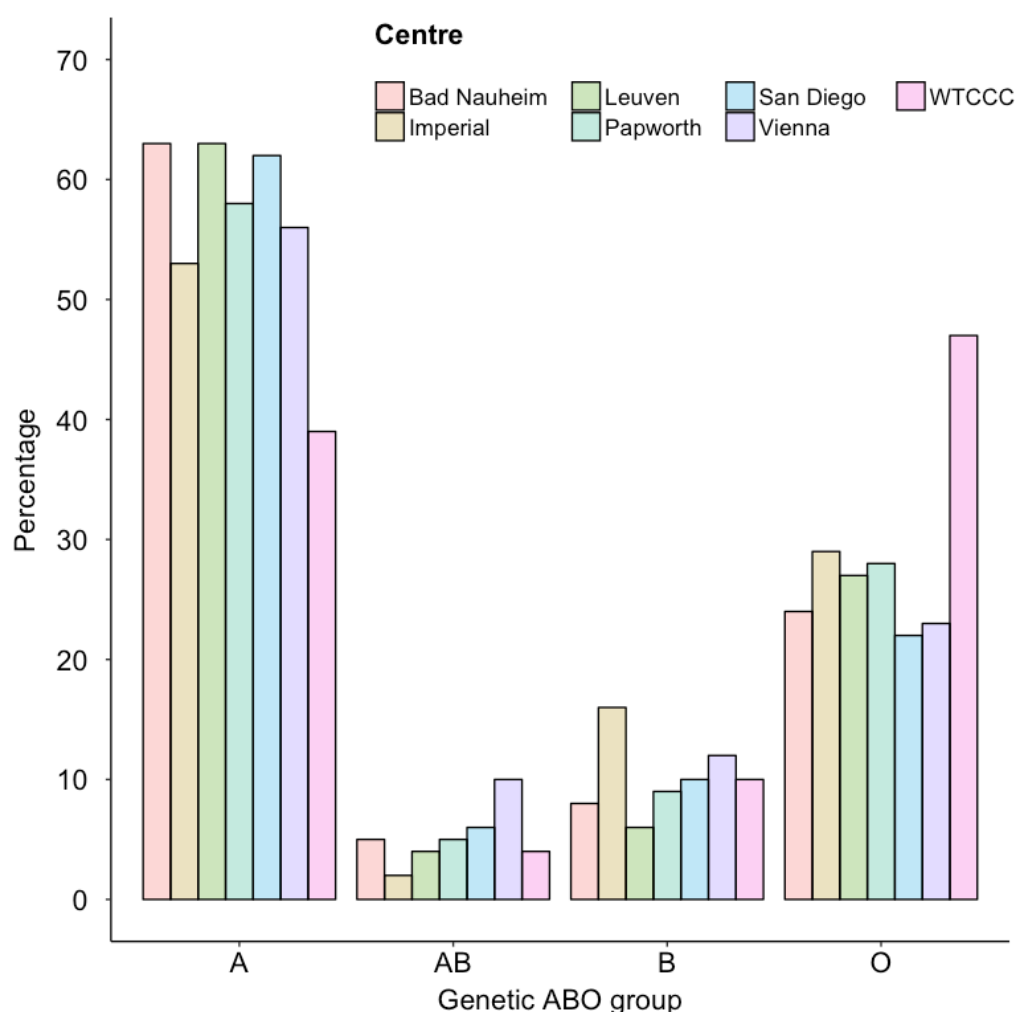


Figure 3.30 The percentage of genetic *ABO* groups in recruiting study centres
WTCCC (Wellcome Trust Case Control Consortium, the healthy control group)

3.2.5 Fine mapping

3.2.5.1 Credible set analysis

A Bayesian analysis was performed for the combined cohort (discovery and validation) and used to calculate a 99% credible SNP set ([Section 2.1.9.1](#)) ([Table 3.11](#) and [Figure 3.31](#)). This comprised the lead SNP (rs2519093) (in the frequentist association testing) and the next two most significant SNPs (rs532436 and rs507666). Rs507666 is the SNP used to tag the A1 genetic *ABO* group described in [Section 3.2.4](#).

There is a high degree of correlation between rs2519093, rs532436 and rs507666 ($R^2=1.00-0.992$, European (non-Finnish) 1000Genomes phase 3 data).

rsID	CHR:POS_EA/NEA	BF	PP	CumPP	Rank
rs2519093	9:136141870_T/C	2.59e+26	4.60e-01	0.460	1
rs532436	9:136149830_A/G	2.03e+26	3.60e-01	0.820	2
rs507666	9:136149399_A/G	1.01e+26	1.79e-01	0.999	3
rs635634	9:136155000_T/C	6.15e+23	1.09e-03	1.000	4
rs600038	9:136151806_C/T	8.52e+21	1.51e-05	1.000	5

Table 3.11 Fine mapping: 99% credible SNP set for the chromosome 9 association

Bayesian analysis was performed in SNPtest as described in [Section 2.1.9.1](#). The posterior probabilities were then calculated by dividing the Bayes factor for each SNP (within a 200kb region of the peak associated SNP) by the sum of all Bayes factors for that region (5.64e+26). Posterior probabilities were ranked in descending order and the SNPs included in the 99% cumulative sum comprised the 99% credible set. Bayes factors and posterior probabilities are displayed using exponential notation. BF (Bayes factor), PP (posterior probability), CumPP (cumulative posterior probability).

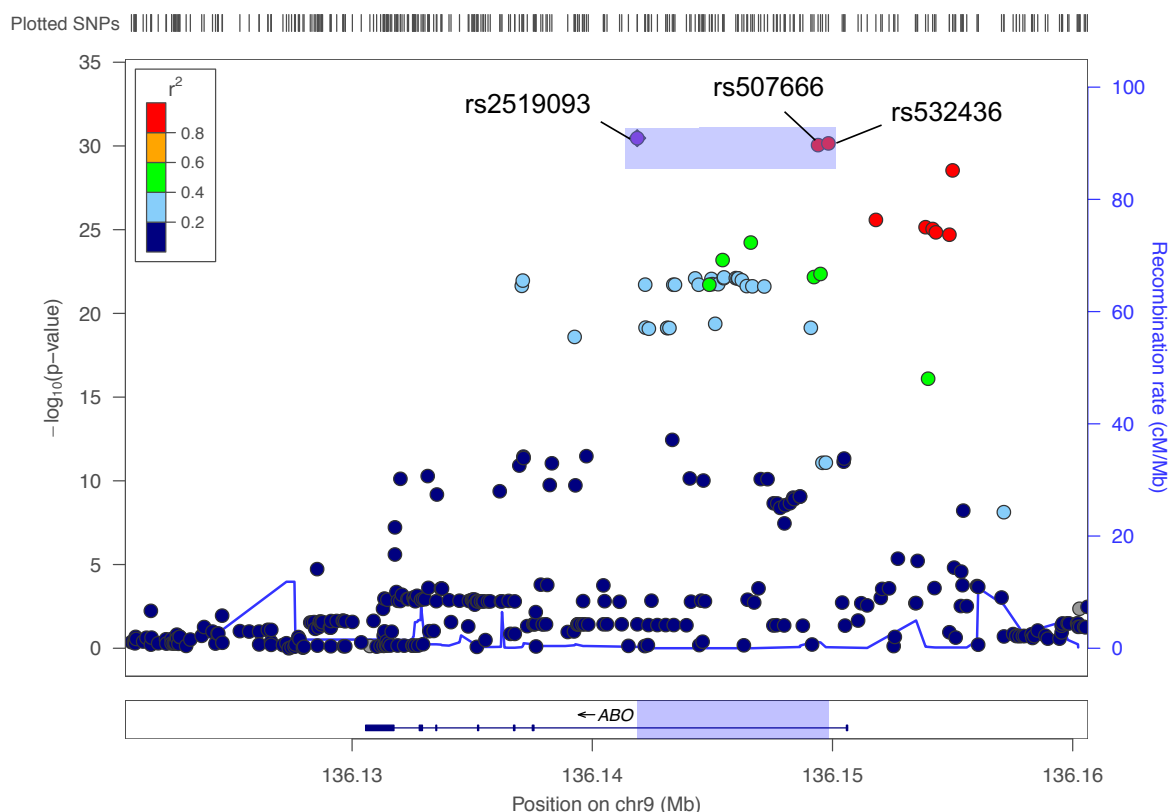


Figure 3.31 Regional association plot of 99% credible SNP set for the chromosome 9 association

The 3 SNPs that comprise the 99% credible set are named and highlighted with the blue rectangle.

3.2.5.2 Genomic functional annotation

The fumaGWAS tool was used to visualise functional annotations for associated loci, using a range of data resources that are described in [Section 2.1.9.2](#). The lead SNP (rs2519093) had a CADD score of 6.85 (unlikely to be highly deleterious), no available evidence of regulatory elements (regulomeDB score=7) and is an eQTL for the *SURF1* gene in the atrial appendage of the heart ($p = 7.02 \times 10^{-8}$, normalised effect size (NES) = -0.380, GTEx v7 data) ([Figure 3.32](#)). rs532436 and rs507666 have CADD scores suggesting low deleteriousness (3.11 and 4.11 respectively) and regulomeDB scores suggesting minimal evidence of regulatory elements (3a and 4 respectively, see [Figure 3.32](#) for explanation). Both rs532436 ($p = 6.95 \times 10^{-8}$, NES = -0.371) and rs507666 ($p = 6.95 \times 10^{-8}$, NES=-0.371) are also eQTLs for the *SURF1* gene in the atrial appendage of the heart.

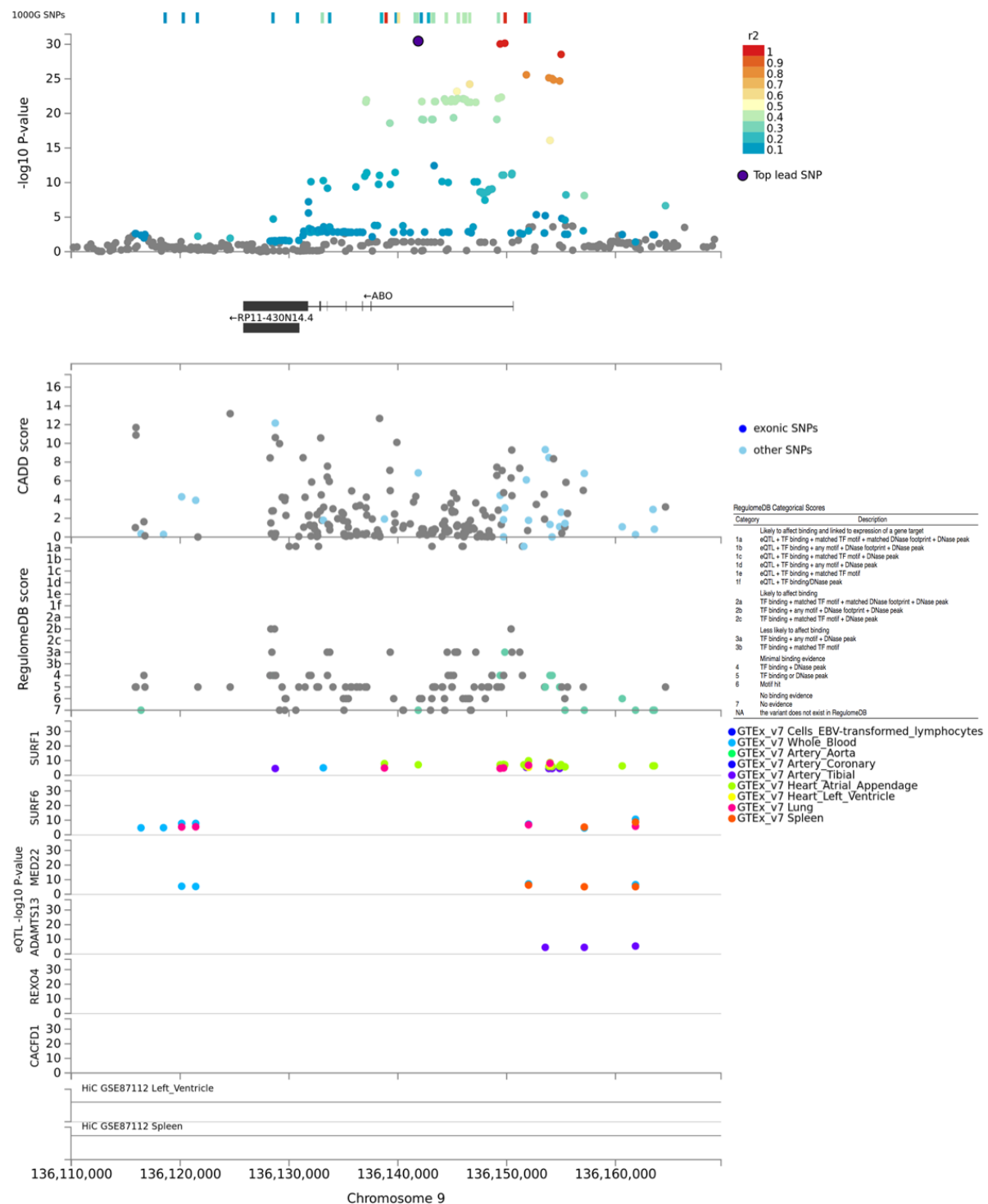


Figure 3.32 Fine mapping: functional annotations for the 99% credible SNP set in chromosome 9

The top panel shows the *ABO* associated locus with correlated SNPs (LD calculated from 1000 Genomes phase 3 data).

The CADD score is shown in the second track with a higher score indicating increased likelihood of deleteriousness.

The third track displays the regulomeDB score (range 1-7) for SNPs, with a lower score indicating greater evidence of regulatory elements.

Cis-expression quantitative trait loci (eQTL) are shown in the fourth panel for selected tissues (GTEx v7: lymphocytes, whole blood, aorta, coronary artery, tibial artery, heart (atrial appendage, left ventricle), liver, lung, spleen). Different colours represent different tissue/cell types. *P*-values on the y-axis are -log₁₀ scale.

There are no chromatin interactions present for genomic region in the selected tissues (aorta, left ventricle, liver, lung, right ventricle, spleen), displayed in the final, empty track (Hi-C data).

SNPs coloured grey are those not in LD ($R^2 < 0.1$) with lead SNP (top panel) or those that were not used for the mapping of the respective tracks.

Additionally, rs507666 was associated with the expression of *SURF1* in left ventricle and lung tissues ($p = 1.56 \times 10^{-5}$, NES = -0.288 and $p = 2.09 \times 10^{-5}$, NES = -0.190, respectively).

There were no chromatin interactions present for the genomic region containing the 99% credible SNP set ([Figure 3.32](#)) in selected tissues (aorta, left ventricle, liver, lung, right ventricle, spleen) using Hi-C (a chromosome conformation capture method) data via fumaGWAS.(255)

There were no additional effects when visualising chromatin interactions and the cis-eQTLs over a wider genomic range (by chromosome) ([Figure 3.33A](#) and [3.33C](#)). The lead SNP (rs2036914) in the chromosome 4 locus (discovery cohort) was an eQTL for the *F11* gene ($p = 3.74$, NES = -0.247). Rs4253409 is moderately correlated ($R^2=0.615$) with the lead chromosome 4 SNP and is an eQTL for the *KLKB1* gene in aortic and tibial artery tissues ([Figure 3.33B](#) and [3.33D](#)).

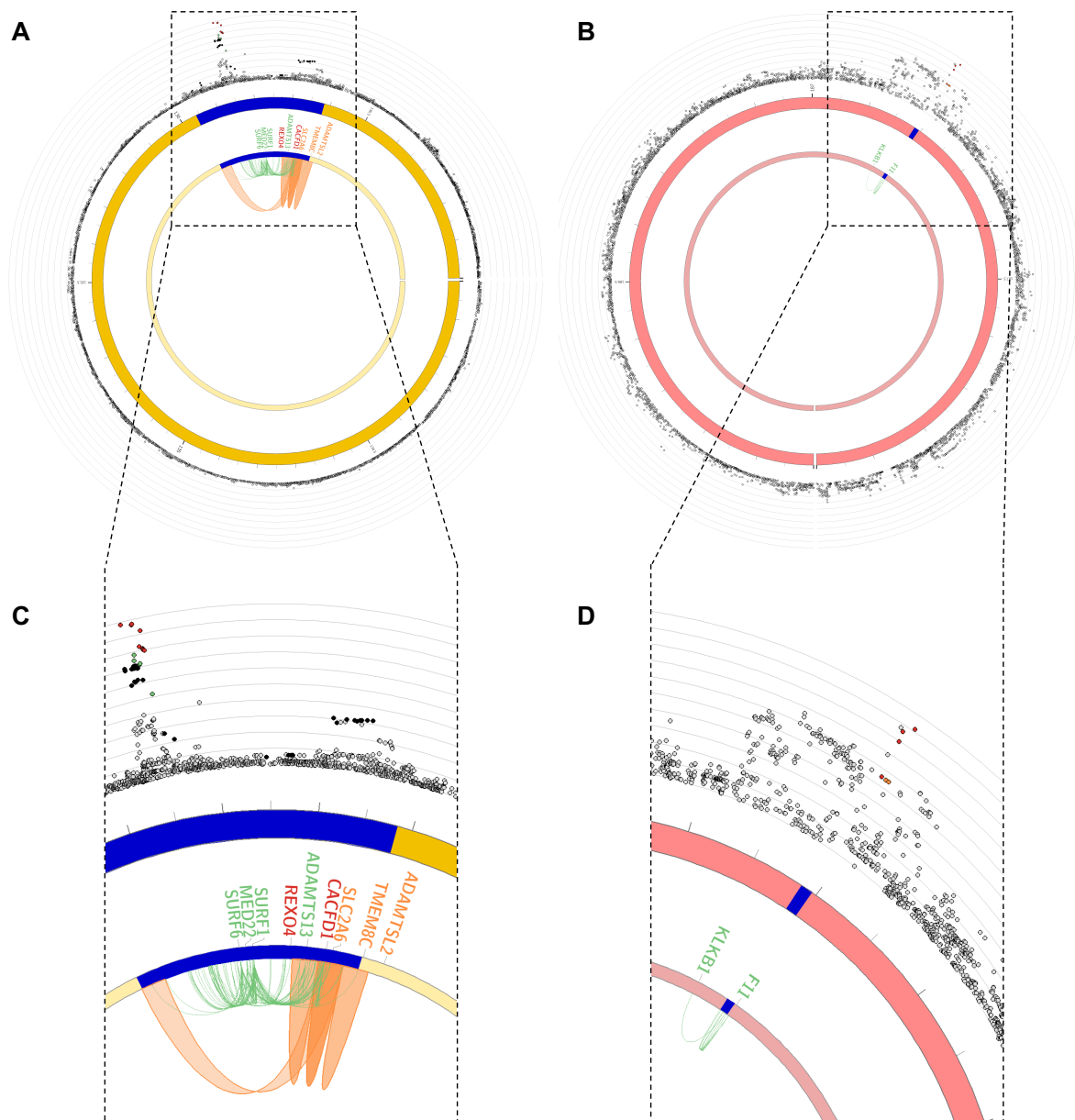


Figure 3.33 Circos plots of chromatin interactions and eQTLs for the associations in chromosome 9 (combined analysis) and chromosome 4 (discovery cohort)

The outer layer displays a Manhattan plot of SNPs and the associated loci in **A** chromosome 9 (combined analysis) and **B** chromosome 4 (discovery cohort). The next layer (yellow or red) is the chromosome co-ordinates with the risk locus in blue. The inner circle displays the eQTLs (green) and chromatin interactions (orange) for the mapped genes. When a gene is mapped to both, it is coloured red. Zoomed in

views of **A** and **B** are shown in **C** and **D** respectively. The chromatin interactions use Hi-C data and the eQTLs are from GTEx v7.(255, 256)

3.2.6 Gene-based and gene-set analysis

Genome-wide association testing was repeated with genes rather than individual SNP markers using MAGMA via fumaGWAS.(226) SNPs were mapped to 19,311 protein coding genes and association testing was performed to assess the joint effect of multiple SNP markers across genes. The *FGG* and *CACFD1* genes were statistically significant across the genome when adjusting for multiple testing (Figure 3.34). Importantly, *ABO* was not included in the genes that were tested as it was unavailable on the panel of genes used by fumaGWAS. The *CACFD1* gene is ~200kb downstream of *ABO* and in close proximity to *ADAMTS13*, which was shown to be in moderate-low LD with *ABO* (Section 3.2.3.1.3). Furthermore, SNPs in the *CACFD1* are associated with CTEPH in the joint analysis (Table 3.6), but this association is not independent when a conditional analysis is performed (Figure 3.23). Consequently, the *CACFD1* gene association is likely to be a proxy association of *ABO*.

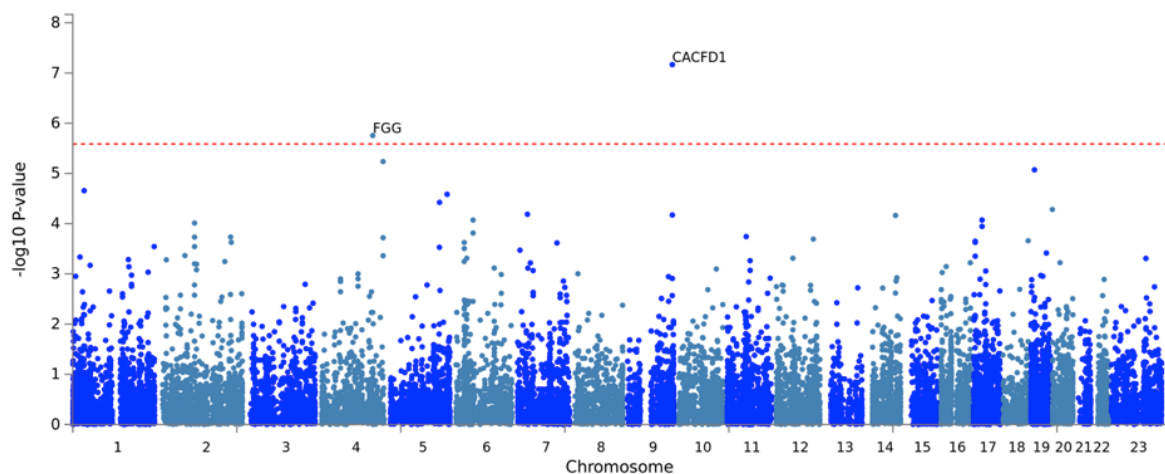


Figure 3.34 Gene-based association testing: combined group (discovery and validation cohort)

Gene-based association was performed using MAGMA via fumaGWAS for 1250 CTEPH patients and 1492 healthy controls. SNPs (n=7,675,738) were mapped to 19,311 protein coding genes. MAGMA performed multiple regression (SNP-wise model) using the summary statistics data from the combined GWAS analysis

described in [Section 3.2.3.1.2](#). The horizontal red line represents a genome-wide significance threshold of $p=2.59 \times 10^{-6}$ (0.05/19,311).

Gene-set analysis was then performed utilising the gene-based p -values for 4,728 curated gene sets and 6,166 gene ontology terms.(221) The most significantly associated gene-set was for cardiac muscle contraction however, this was not statistically significant when adjusted for multiple testing ([Table 3.12](#)). The most associated tissue in the MAGMA tissue expression analysis was whole blood, followed by atrial appendage and left ventricle of the heart ([Figure 3.35](#)). No tissue was significantly associated after adjustment for multiple testing.

Gene Set	N genes	Beta	P	P_{bon}
Curated_gene_sets: kegg_cardiac_muscle_contraction	72	0.364	6.08e-5	0.647
GO_bp: go_cytoplasmic_translation	39	0.452	1.12e-4	1.000
Curated_gene_sets: korkola_embryonal_carcinoma_dn	12	0.877	1.78e-4	1.000
Curated_gene_sets: biocarta_intrinsic_pathway	23	0.581	2.18e-4	1.000
GO_bp: go_regulation_of_amine_transport	71	0.339	2.34e-4	1.000
Curated_gene_sets: creighton_endocrine_therapy_resistance_5	460	0.135	3.04e-4	1.000
Curated_gene_sets: missiaglia_regulated_by_methylation_up	117	0.256	4.33e-4	1.000
Curated_gene_sets: korkola_seminoma_dn	11	0.827	4.73e-4	1.000
Curated_gene_sets: kondo_colon_cancer_hcp_with_h3k27me1	26	0.53	5.24e-4	1.000
Curated_gene_sets: matzuk_meiotic_and_dna_repair	38	0.386	6.21e-4	1.000

Table 3.12 MAGMA gene-set analysis

MAGMA gene-set analysis was performed via fumaGWAS. P -values from the gene-based analysis were utilised for an analysis with curated gene sets and GO terms obtained from MsigDB. Only the top 10 most significant gene-sets are displayed. No gene-set was significantly associated when adjusted for multiple testing. N genes (number of genes in the gene-set), P_{bon} (p -value adjusted for Bonferroni correction).

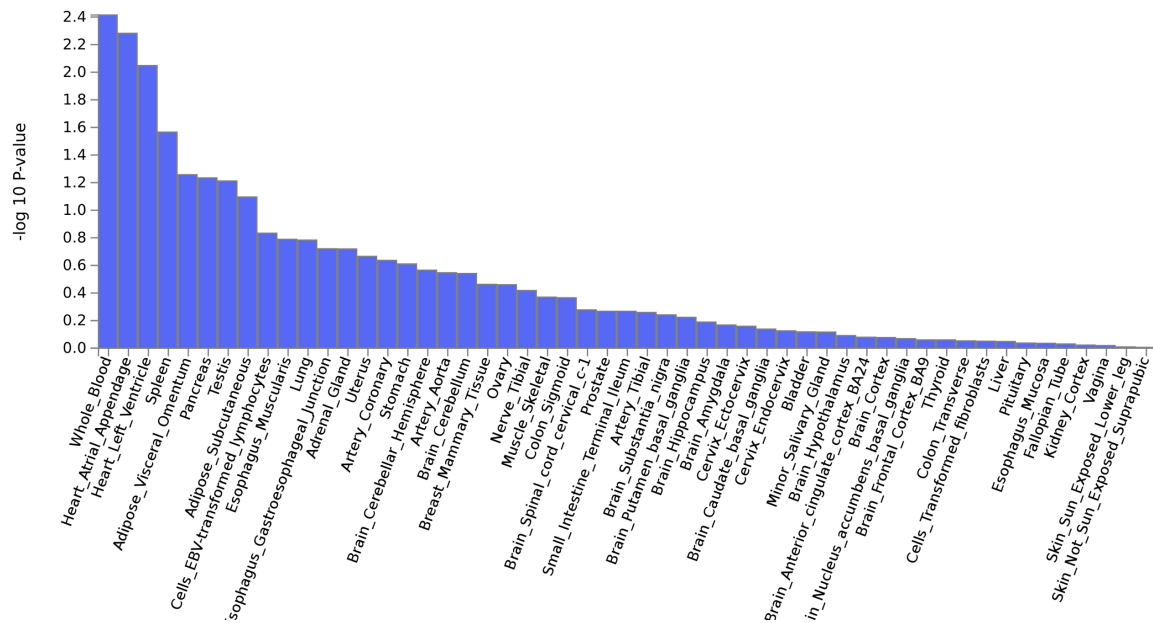


Figure 3.35 MAGMA tissue expression analysis

MAGMA tissue expression analysis was performed using the p -values from the gene-based analysis and tissue specific gene expression values from GTEx v7 for 53 general tissue types. Average gene expression values are log2 transformed.(221) No tissues were significantly associated when adjusted for false discovery rate.

3.2.7 GWAS putative associations

To assess putative associations, loci with a higher p -value threshold ($p < 1 \times 10^{-5}$) are summarised in [Table 3.13](#). These regions may contain genuine SNP-trait associations that are currently under-powered with current sample numbers. Care should be taken in interpreting the putative associations, as lowering the p -value threshold also increases false positive associations.

rsID	CHR:POS_EA/NEA	GENE	FUNC	EAF_A	EAF_U	EAF_REF	INFO	OR (95% CI)	p
rs7594443	2:206668450_C/T	NRP2	intergenic	0.027	0.047	0.037	0.766	2.33 (1.97-2.69)	3.71e-06
rs6794945	3:133518463_C/T	SRPRB	intronic	0.504	0.550	0.666	0.695	1.68 (1.47-1.89)	1.25e-06
rs13130318	4:155538470_G/T	FGG	intergenic	0.318	0.241	0.219	0.935	1.41 (1.28-1.54)	4.45e-07
rs264994	5:79165176_T/C	CMYA5 / MTX3	intergenic	0.026	0.046	0.067	0.730	2.43 (2.06-2.80)	2.49e-06
rs678409	6:74299575_C/T	SLC17A5	intergenic	0.601	0.655	0.624	0.928	1.34 (1.22-1.46)	3.79e-06
rs117853706	9:6972010_G/A	KDM4C	intronic	0.989	0.980	0.968	0.809	3.62 (3.05-4.19)	9.80e-06
rs72784778	10:31961058_T/A	RP11-472N13.3	intergenic	0.970	0.953	0.975	0.916	2.17 (1.83-2.50)	5.79e-06
rs12413249	10:6714659_C/T	RP11-554I8.1	intergenic	0.783	0.831	0.801	0.992	1.41 (1.25-1.56)	9.59e-06
rs7297105	12:127495362_C/T	RP11-575F12.1	intronic	0.411	0.468	0.430	0.719	1.42 (1.28-1.56)	1.02e-06
rs73324509	14:90404239_T/G	EFCAB11	intronic	0.014	0.031	0.026	0.944	2.72 (2.27-3.17)	9.71e-06
rs142103720	18:33140192_G/A	GALNT1	intergenic	0.984	0.969	0.975	0.780	2.74 (2.30-3.18)	8.45e-06

Table 3.13 Putative GWAS associations: joint analysis group

Putative associations ($p < 1 \times 10^{-5}$) from the analysis described in [Section 3.2.3.1.2](#) (1250 CTEPH patients, 1492 healthy controls and 7,675,738 SNPs). The associations previously described in chromosome 9 (*ABO*) and 4 (*F11*) are not included in the table. Only the lead (most significant) SNP is shown. The column headings and additional details are described in [Table 3.6](#). Gene names starting with “RP11” are long non-coding RNAs.

3.3 Discussion

3.3.1 Overview

This multi-centre international GWAS is the largest study undertaken in CTEPH. The *ABO* locus was identified as the most significant common variant genetic association in CTEPH in both a discovery and validation cohort. There was a putative association in the discovery cohort in the *F11* gene locus. The A1 subgroup of *ABO* was enriched in CTEPH and this may result in multiple functional consequences given the pleiotropy displayed by *ABO*.

3.3.2 Associated Loci

3.3.2.1 The *ABO* association

The most significant SNP associated with CTEPH is an intronic variant in the *ABO* gene (rs2519093, OR (95% CI) = 2.4 (2.3-2.5), $p=3.42 \times 10^{-31}$). The effect allele is over-represented in CTEPH cases compared with healthy controls in the current study and a reference European population (0.316, 0.165 and 0.182 respectively). A conditional analysis did not identify any additional variants that were independently associated with CTEPH at the chromosome 9 locus. Fine mapping using a Bayesian analysis identified a 99% credible set of 3 SNPs (rs2519093, rs507666 and rs532436), which are most likely to contain the causative variant. The limitation of this analysis is that it assumes the causative SNP is present in the region tested and that only one variant is causative.(257) One of these SNPs (rs507666) is the variant used to “tag” the A1 genetic *ABO* group.

When genetic *ABO* subgroups were reconstructed, the risk of CTEPH was increased for the A1 enriched groups, which was most marked in the A1A1 group (OR (95% CI) 4.39 (2.92-6.69), $p<0.001$). CTEPH risk was also increased to a lesser degree in group B individuals but the A2 group was not associated. This is consistent with VTE, which also exhibits an association with A1 enriched *ABO* subtypes but not with the A2 group. (253) Individuals possessing A1 enriched *ABO* groups have higher plasma levels of VWF and factor VIII compared with the O and A2 enriched groups.(253) This functional consequence of *ABO* may be an aetiological mechanism in CTEPH.

The ABO groups vary by the ABH(O) antigens (oligosaccharide residues) and are found on red blood cells, platelets and VWF, a protein involved in haemostasis.(80) Genetic variation in *ABO* has been associated with ischaemic stroke, coronary artery disease and venous thromboembolism.(70, 258) Whilst the exact mechanism linking ABO antigen groups to thrombotic risk has not been defined, it may be mediated by VWF levels, which are 25% lower in O group individuals.(121) The relationship between ABO groups and the ADAMTS13-VWF axis is discussed further in [Chapter 4](#).

ABO is likely to have additional functional effects as it is a pleiotropic locus that is associated with a large number of diseases and traits. There are 99 studies, 236 associations and 78 traits related to the *ABO* locus recorded in the GWAS Catalog (accessed March 2019).(140) The *ABO*-disease associations include: VTE (259), CAD (260), ischaemic stroke (261), allergy (262) and type 2 diabetes mellitus (263). Additional traits associated with *ABO* include: coagulation factor levels (VWF and factor VIII) (264), alkaline phosphatase (265), blood cell trait variation (266), P- and E-selectins (267, 268), lipid traits (269), endothelial growth factors (270) and Intercellular Adhesion Molecule 1 (ICAM-1) (271). The GWAS Catalog traits associated with the 3 SNPs comprising the CTEPH GWAS 99% credible set (rs2519093, rs507666 and rs532436) include VTE, CAD, ICAM-1, and lipid levels.(140)

A recent study charting the human plasma proteome identified 64 plasma protein levels that were associated with *ABO* variants. (272) Of these associations, there were 12 in the 3 SNPs (rs2519093, rs532436 and rs507666) that comprised the 99% credible set in the CTEPH GWAS. The protein levels influenced by these *ABO* variants included: P- and E-selectins, Interleukin-3 receptor subunit alpha (IL3RA), Vascular endothelial growth factor receptor 3 (FLT4), Protein FAM3B (FAM3B), Thrombospondin type-1 domain-containing protein 1 (THSD1), Insulin receptor (INSR) and ICAM-1.(272) *ABO* is a pleiotropic locus that may have a wide range of functional consequences that result in thrombotic disease including CTEPH.

3.3.2.2 The *F11* putative association

There was a significant association in the discovery cohort with a *F11* gene intronic variant (rs2036914, OR (95% CI) = 1.43 (1.30-1.56), $p = 4.79 \times 10^{-8}$) that was not

replicated in the validation cohort. This SNP has also been associated with venous thromboembolism and activated partial thromboplastin time.(70, 273)

Coagulation factor XI (FXI) is a component of the blood coagulation pathway that acts downstream of factor XII, and is able to activate factors FX, FV and FVIII.(274) FXI has a significant role in promoting thrombosis and is associated with ischaemic stroke and myocardial infarction in addition to VTE.(274)

3.3.3 Genomic functional annotations, gene-set and gene-based analyses

Genomic functional annotations were utilised to investigate the lead *ABO* SNP (rs2519093). This variant is an expression quantitative trait locus (eQTL) for the Surfeit locus protein 1 (*SURF1*) gene on chromosome 9 particularly in the atrial appendage of the heart. *SURF1* is associated with oxidative phosphorylation, which has been implicated in the pathobiology of other forms of PH.(275) Right ventricular adaptation is important in CTEPH and therefore alterations in the oxidative phosphorylation pathways of the heart is a plausible pathobiological mechanism in CTEPH.(276)

Exploratory gene-set and gene-based analysis was performed as an alternative to single variant association testing due to the studies small sample size and relative lack of power. Gene-based analysis identified significant associations in the Fibrinogen Gamma (*FGG*) and calcium channel flower domain containing 1 (*CACFD1*) genes. The *CACFD1* association was unlikely to have been independent from *ABO* and therefore not unique. *FGG* has been associated with venous thromboembolism and is discussed in [Section 3.3.4.1](#). There were no significant associations for the gene-set and tissue expression analyses.

3.3.4 Absence of genetic associations

As well as the presence of genetic associations, the absence of associations in the CTEPH GWAS may be also informative. SNPs in *FGG* and genetic variants that occur in other types of pulmonary hypertension which have previously been investigated in CTEPH are discussed in [Sections 3.3.4.1](#) and [3.3.4.2](#). Genetic variants in the CTEPH GWAS that relate to previously described VTE associated variants and warfarin metabolism associations are investigated in [Chapter 5](#).

3.3.4.1 The *FGA-FGB-FGG* locus

No single variant in the *FGA-FGB-FGG* locus was associated with CTEPH in the GWAS. A SNP in the *FGA* gene (rs6050; missense variant) encoding the fibrinogen A α chain protein has previously been associated with CTEPH.(42, 66) *Suntharalingam et al*, described that the rs6050 polymorphism was over-represented in 214 CTEPH cases compared with 200 healthy controls.(66) A subsequent study by *Li et al* in 101 patients with CTEPH, 102 with pulmonary embolism and 108 healthy controls confirmed that rs6050 was overrepresented in CTEPH compared with healthy controls but not in PE.(42) Furthermore, fibrin resistance has been demonstrated to vary in CTEPH patients with rs6050 genotype and this SNP variant is associated with clot structure that may predispose to embolisation.(42, 277) These CTEPH studies were limited by small sample sizes (for genetic studies) and an inability to adjust for important potential confounding factors including population structure.

Rs6050 was poorly imputed in the current study and therefore not included in the CTEPH GWAS analysis. However, rs6050 is highly correlated (LDlink: $R^2 = 0.872$, $p < 0.001$, European (non-Finnish)) with rs13130318 (OR (95% CI) = 1.41 (1.28-1.54), $p = 4.45 \times 10^{-7}$) a putative CTEPH association ([Table 3.13](#)). Furthermore, rs6050 is highly correlated with rs2066865 ($R^2 = 0.879$) which is associated with VTE in a GWAS meta-analysis ([Chapter 5, Table 5.2](#)). The absence of an *FGA-FGB-FGG* locus association using single variant GWAS statistical testing may reflect a lack of study power. There is putative evidence of an association between CTEPH and the *FGA-FGB-FGG* locus when a higher p -value threshold ($p < 1 \times 10^{-5}$) is applied or when utilising gene-based analysis. Gene-based analysis has the advantage that less statistical tests are performed by grouping variants into genes which can increase power to detect associations compared with single-variant analysis.(226) They are limited by assigning non-coding variants to adjacent genes, which may not be the causal variant/gene in some circumstances.(278) Nevertheless, the putative association in the single-variant GWAS analysis and the gene-based significance of the *FGA-FGB-FGG* suggest that this is a likely association that may become significant as sample size increases.

3.3.4.2 Pulmonary hypertension related genes in CTEPH

Genetic variants that occur in other types of pulmonary hypertension have been previously investigated in CTEPH and are summarised in [Section 1.3.3](#). In the CTEPH GWAS, there is no association with common variants in these PH related genes. However, many of the PH associated variants are rare / very rare and would not be included in this GWAS analysis. Previous CTEPH studies that have reported genetic variant associations have used a candidate gene / variant approach and methodologies that may have resulted in false positive associations.(131) Alternatively, some genetic variants associated with CTEPH may be specific to the population being studied and their absence in the current study may be a consequence of the Caucasian composition.

3.3.5 Strengths and limitations

This multi-centre international GWAS is the largest study undertaken in CTEPH, an uncommon disease that occurs following acute pulmonary embolism. CTEPH patients included in the study had a robust diagnosis that is defined at expert PH centres using a range of investigations and international guidelines. The significant *ABO* locus occurs in VTE and is therefore likely to represent a genuine association.

Separate studies from *Bonderman et al* have shown a prevalence of non-O ABO blood groups of 77% in CTEPH patients compared with 58% in non-CTEPH and an odds ratio 2.1 (95% CI 1.1-3.9) for the non-O blood group.(12, 35) In an international CTEPH registry of 679 patients, the prevalence of non-O blood groups was 76% and this compares with 55% in a UK biobank cohort.(68, 279) In the current GWAS, inferred non-O blood groups occurred in 74% of CTEPH patients compared with 43% in healthy controls. The current GWAS finding of an association between *ABO* and CTEPH is more robust than previous observational studies that had not adjusted adequately for confounding factors particularly population structure, which is important as ABO frequencies vary widely with ethnicity.(280)

Whilst *ABO* is definitively associated with CTEPH in this study, a limitation of the GWAS methodology is that it does not necessarily prove aetiological causality or provide a mechanism by which genetic *ABO* association predisposes to CTEPH. Alternatively, patients with CTEPH may have *ABO* associations due to another

“hidden” confounder that causes CTEPH rather than directly via *ABO*, or *ABO* may interact with other genes/variants predisposing to CTEPH, which is possible given the pleiotropic nature of the *ABO* locus. The GWAS study cohort is biased towards containing patients with more severe CTEPH that are being assessed for PEA. Therefore, *ABO* may be associated with CTEPH disease severity and progression rather than disease aetiology, which is addressed in [Chapter 5](#).

The current study sample size limits the ability to identify genetic variants associated with CTEPH with more modest effect sizes. This may account for the absence of previously reported VTE genetic associations that may be expected given that three-quarters of CTEPH patients have had a preceding VTE. This is investigated in [Chapter 5](#). There were a large number of samples (CTEPH 305, health controls 44) that were removed during the quality control steps. This was necessary to account for variable genotyping quality between batches and to adhere to robust GWAS methodology. Alternative statistical approaches such as linear mixed modelling could be utilising to retain some samples (e.g. those removed due to ethnicity), although this is likely to have minimal impact on the overall study power.⁽²⁸¹⁾ Shared control samples were used between the discovery and validation cohort and whilst this is an accepted methodology, power could have been increased by having separate control groups.^(204, 282) Furthermore, comparing allele frequencies from the validation cohort that contained European and American patients to a UK based control group required additional statistical correction for population structure and it is possible a degree of residual confounding remains. Ideally a population matched control group should be used for the discovery and validation cohorts, although this can be challenging in modern international GWAS consortia.

The fine mapping using genomic functional annotations was limited by the tissue types available in the reference datasets. Non-coding variants can affect transcriptional regulation differently dependent upon the tissue and cell type.⁽¹⁵⁶⁾ Therefore, annotations should ideally be interrogated in the tissues implicated in disease pathobiology, which for CTEPH would primarily be pulmonary vascular endothelial cells, right heart samples and blood cells. The most complete set of tissue types that were available were analysed from annotation datasets, including 53 from GTEx and

127 from ENCODE, however neither contained pulmonary vascular endothelial cells or right heart tissue which limited the analysis.(221)

In summary, the *ABO* locus is associated with CTEPH in a GWAS and this is driven by an over-representation of the A1 subtype. The genetic *ABO* association may result in functional consequences related to CTEPH pathobiology.

4 The ADAMTS13-VWF axis

4.1 Introduction

The *ABO* gene locus was the most significant association in the CTEPH GWAS ([Chapter 3](#)). The *ADAMTS13* gene locus is situated ~200kb distal of *ABO* and in low-moderate linkage disequilibrium with *ABO*. Whilst the *ADAMTS13* locus was initially associated in the GWAS, this was not independent of the *ABO* association.

The *ABO* gene is linked to the ADAMTS13-VWF axis, as *ABO* groups determine a large proportion of the variation in VWF.(200) ADAMTS13 has only one known substrate, VWF and therefore, it regulates VWF activity by cleaving the more procoagulant ultra-large VWF multimers.

Increased VWF and reduced ADAMTS13 are associated with thrombotic diseases including coronary artery disease, ischaemic stroke and venous thromboembolism.(191-193) Moreover, micro-vessel thrombosis occurs when ADAMTS13 activity levels are severely decreased by autoantibodies in thrombotic thrombocytopenic purpura.(188)

Abnormalities in haemostasis including elevated VWF and an association with non-O blood groups have been described in CTEPH.(35, 68) The role of ADAMTS13 in CTEPH has not been investigated to date.

The aim of this Chapter was to investigate the ADAMTS13-VWF axis in CTEPH patients including its relationship to *ABO* groups and *ADAMTS13* genetic variants.

4.2 Results

4.2.1 Study samples and participants

ADAMTS13 and VWF plasma concentrations were measured in 208 CTEPH patients and 68 healthy controls. Additional disease groups comprised: 35 patients with CTED, 30 with IPAH and 28 following PE. Baseline group characteristics are summarised in [Table 4.1](#) and [Table 4.2](#). Age and sex differed across the groups ($p<0.001$ and $p=0.014$) with CTEPH patients being older (median \pm IQR: 64 ± 19 years) than healthy controls (49 ± 24 years). As ADAMTS13 levels can vary with age and sex, multivariable linear regression was used to adjust for these variables ([Section 4.2.2.1](#)).⁽²⁰¹⁾ Ethnicity also differed ($p<0.001$) with more non-Caucasians in the PE group. In the whole CTEPH group, 176 (87%) had a proximal distribution of pulmonary arterial obstruction deemed to be surgically accessible and 150 (72%) underwent PEA.

	Healthy control	CTEPH	CTED	IPAH	PE	<i>p</i>
Subjects	68	208	35	30	28	<0.001
Age, Years	49 ± 24	64 ± 19	58 ± 27	64 ± 27	52 ± 26	0.013
Sex, Female	32 (47)	90 (43)	9 (26)	21 (70)	15 (54)	<0.001
Ethnicity, Caucasian	53 (78)	180 (95)	28 (88)	26 (90)	13 (54)	<0.001
WHO functional class						
1		4 (2)	6 (18)	5 (17)		<0.001
2		42 (21)	17 (50)	4 (13)		
3		151 (74)	11 (32)	21 (70)		
4		7 (3)	0 (0)	0 (0)		
6mwd, Metres		318 ± 176	366 ± 180	342 ± 244		0.204
Pulmonary haemodynamics						
mPAP, mmHg		42 ± 18	21 ± 4	42 ± 17		<0.001
CI, L/min/m ²		2 ± 0.6	2.4 ± 0.6	1.7 ± 0.8		<0.001
PVR, dynes.s.cm ⁻⁵		639 ± 476	151 ± 71	808 ± 642		<0.001
Clinical blood tests						
Haemoglobin, g/dL		140 ± 27	138 ± 16	142 ± 22		0.848

Platelet count, x10 ⁹		246 ± 82	200 ± 56	222 ± 77		0.014
WCC, x10 ⁹		7 ± 3	6.6 ± 2.1	6.9 ± 2.4		0.273
Lymphocyte, %		25 ± 10	28 ± 13	18 ± 13		0.025
Neutrophil, %		64 ± 14	59 ± 14	72 ± 14		0.007
CRP, mg/L		5 ± 10	3 ± 3	3 ± 4		0.035
NT-proBNP, pg/mL		592 ± 1576	113 ± 194	334 ± 695		0.006
Smoking status						
Never		91 (47)	16 (50)	15 (52)		0.943
Ex-smoker		87 (45)	13 (41)	11 (38)		
Current smoker		15 (8)	3 (9)	3 (10)		
Anticoagulation medication		137 (94)	15 (94)	30 (100)		0.004

Table 4.1 Baseline group characteristics

Data is presented as median ± interquartile range or number of patients (%). Percentages were calculated using the number of patients that data was available for as the denominator. The differences in categorical variables between groups were assessed using Chi-squared or Fisher's exact test, and the Cochran-Armitage test for WHO functional class. The difference across groups in continuous variables was assessed using the Kruskal-Wallis test; the non-parametric equivalent of a one way analysis of variance (ANOVA). *P*-values adjusted for the number of statistical tests performed using FDR correction. 6mwd (6-minute walk distance), CI (cardiac index), CRP (C-reactive protein) mPAP (mean pulmonary artery pressure), NT-proBNP (N-terminal pro b-type natriuretic peptide), PVR (pulmonary vascular resistance), WCC (white cell count).

	n (%)
CTEPH	
Disease distribution	
Proximal	176 (87)
Distal	25 (13)
PEA	150 (72)
Residual PH (>25mmHg)	83 (63)
Co-morbidities	
IHD	20 (14)
DM	19 (13)
Malignancy	19 (13)
Thrombophilia	9 (6)
Splenectomy	9 (6)
Systemic hypertension	48 (34)
Atrial fibrillation / flutter	14 (10)
COPD	8 (6)
PE	
VQ defects post PE	
None	8 (40)
Present	12 (60)
Idiopathic PE	8 (40)

Table 4.2 Additional clinical phenotype data for the CTEPH and PE groups

COPD (chronic obstructive pulmonary disease), IHD (ischaemic heart disease), DM (diabetes mellitus), PH (pulmonary hypertension), VQ (ventilation-perfusion).

4.2.2 ADAMTS13 and VWF plasma concentrations

4.2.2.1 ADAMTS13 plasma concentrations

ADAMTS13 antigen levels were decreased in CTEPH patients ($0.889 \pm 0.397 \mu\text{g/mL}$; $p < 0.001$) compared to healthy controls ($1.15 \pm 0.300 \mu\text{g/mL}$) (Figure 4.1A). Furthermore, ADAMTS13 was reduced in CTED cases ($0.831 \pm 0.224 \mu\text{g/mL}$, $p < 0.001$) but levels were similar to CTEPH ($p = 0.205$) (Table 4.3). There was no difference in ADAMTS13 levels between IPAH ($1.12 \pm 0.413 \mu\text{g/mL}$; $p = 0.373$) and healthy controls, though the PE group did exhibit slightly lower levels (0.969 ± 0.704 ; $p = 0.049$) (Table 4.3).

Since ADAMTS13 levels can vary with age and sex, group associations were assessed with multivariable linear regression adjusted for age, sex and additionally batch and ethnicity (Table 4.4). This confirmed that ADAMTS13 was lowest in the CTEPH (β (% change) = -23.4%, $p < 0.001$) and CTED groups ($\beta = -25.9\%$, $p < 0.001$). These observations should be interpreted together with the additional models that utilise interaction terms presented in Section 4.2.2.3. Furthermore, increasing age was also associated with lower ADAMTS13 ($\beta = -5\%$ per 10 years, $p < 0.001$). ADAMTS13 antigen levels were not significantly associated with the PE group ($\beta = -12\%$, $p = 0.089$), nor were they associated with IPAH, sex or ethnicity.

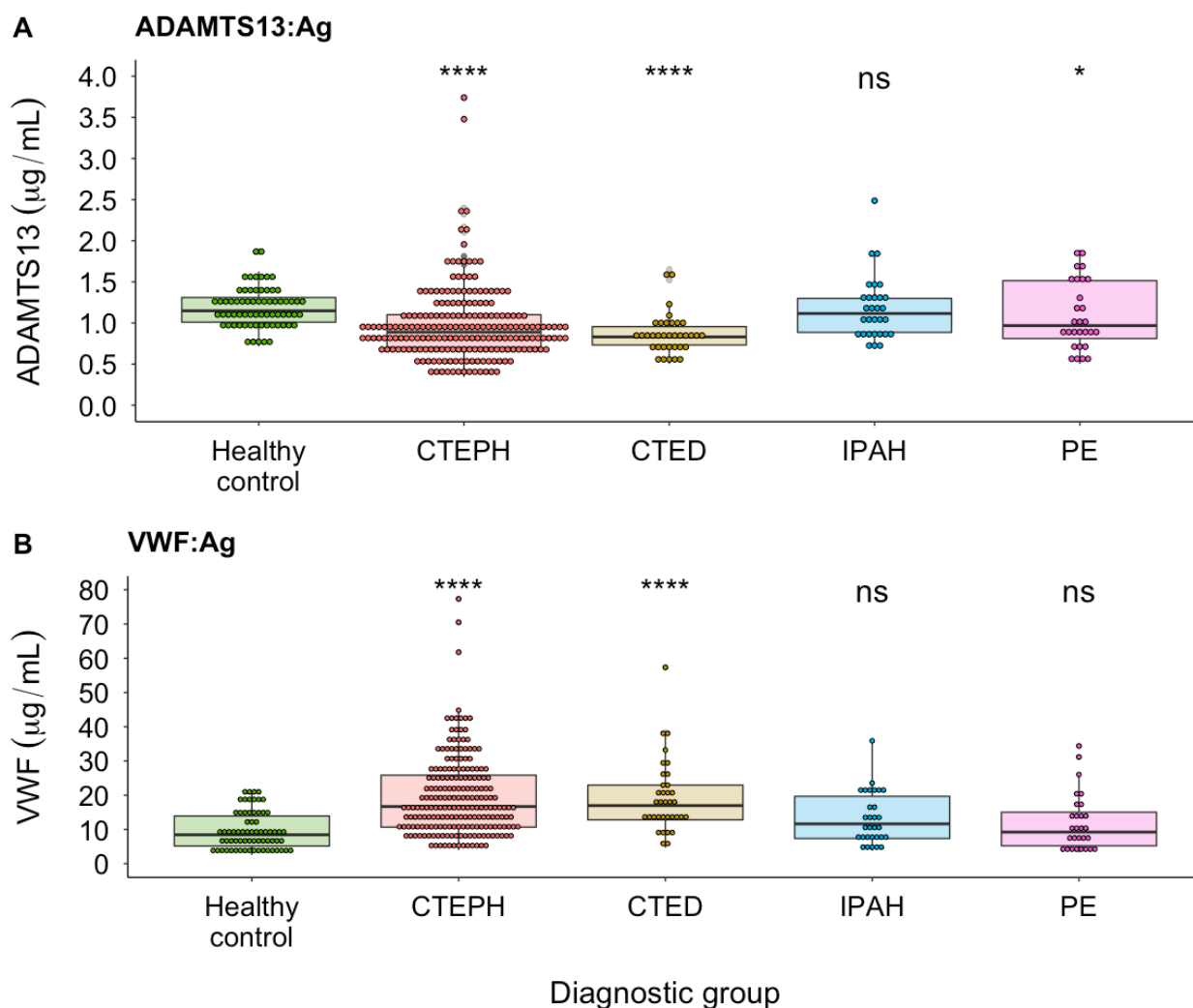


Figure 4.1 ADAMTS13 and VWF antigen (Ag) levels by diagnostic groups

Dunn's test with FDR adjustment was used to calculate p -values. ****: $p \leq 0.0001$, ***: $p \leq 0.001$, **: $p \leq 0.01$, *: $p \leq 0.05$, ns (not significant): $p > 0.05$. Healthy control ($n=68$), CTEPH (chronic thromboembolic pulmonary hypertension, $n=208$), CTED (chronic thromboembolic disease, $n=35$), IPAH (idiopathic pulmonary arterial hypertension, $n=28$), PE (pulmonary embolism, $n=28$).

ADAMTS13				
	Healthy control	CTEPH	CTED	IPAH
CTEPH	3.00×10^{-08}	-	-	-
CTED	9.70×10^{-07}	0.205	-	-
IPAH	0.373	0.003	0.001	-
PE	0.049	0.131	0.038	0.294
VWF				
	Healthy control	CTEPH	CTED	IPAH
CTEPH	4.00×10^{-12}	-	-	-
CTED	2.20×10^{-06}	0.834	-	-
IPAH	0.071	0.006	0.021	-
PE	0.433	1.90×10^{-04}	0.002	0.433

Table 4.3 ADAMTS13 and VWF antigen level pair-wise diagnostic group comparisons

Dunn's test with FDR adjustment was used to calculate *p*-values.

	Model1			Model2*		
	β (%)	95% CI (%)	<i>p</i>	β (%)	95% CI (%)	<i>p</i>
Healthy Control	Reference					
CTEPH	-23.4	-30.9, -15.1	5.91×10^{-07}	-17.6	-27.9, -5.78	0.005
CTED	-25.9	-35.1, -15.4	1.18×10^{-05}	-23.1	-35.1, -8.75	0.003
IPAH	-2.18	-14.7, 12.2	0.752	1.17	-14.7, 20.0	0.894
PE	-12.0	-24.0, 1.97	0.089	-7.84	-23.3, 10.7	0.381
Female	Reference					
Male	-1.07	-7.57, 5.89	0.756	-1.98	-9.92, 6.65	0.641
Age	-0.518	-0.732, -0.303	3.30×10^{-06}	-0.541	-0.810, -0.271	9.99×10^{-5}
Batch1	Reference					
Batch2	-2.16	-10.5, 6.95	0.630	13.5	1.71, 26.8	0.024
Caucasian	Reference					
Non-Caucasian	-5.58	-15.2, 5.10	0.293	-6.7	-18.4, 6.59	0.306

Table 4.4 Multivariable linear regression model of ADAMTS13 antigen levels

Beta (β) coefficients and 95% confidence intervals (95% CI) are presented as percentage change with respect to healthy controls. The reference diagnostic group is healthy control, the reference sex is female, the reference batch is batch1 and the reference ethnicity is Caucasian. n=343 individuals included in the models.

*Model2 additionally adjusted for VWF antigen levels.

4.2.2.2 VWF plasma concentrations

VWF antigen levels were confirmed to be increased in CTEPH (16.7 ± 15.2 $\mu\text{g/mL}$; $p < 0.001$) compared to healthy controls (8.45 ± 8.77 $\mu\text{g/mL}$) and PE (9.23 ± 9.82 $\mu\text{g/mL}$; $p < 0.001$) ([Figure 4.1B](#)). Furthermore, VWF was increased in CTED (17.0 ± 10.1 $\mu\text{g/mL}$, $p < 0.001$) compared to healthy controls, but were no different to CTEPH ($p = 0.834$) ([Table 4.3](#)). There was no difference in VWF antigen levels between IPAH (11.6 ± 12.3 $\mu\text{g/mL}$; $p = 0.071$) or PE ($p = 0.433$) and healthy controls.

Multivariable linear regression was also used for VWF plasma concentrations as described for ADAMTS13. This confirmed that VWF was significantly increased in the CTEPH ($\beta=+75.5\%$, $p<0.001$) and CTED groups ($\beta=+89.5\%$, $p<0.001$) (Table 4.5). Furthermore, increasing age was also associated with increased VWF ($\beta=+6\%$ per 10 years, $p=0.005$). VWF plasma concentrations were not significantly associated with the IPAH or PE groups, sex or ethnicity.

	β (%)	95% CI (%)	p
CTEPH	75.5	44.8, 113	2.00×10^{-8}
CTED	89.5	48.0, 143	6.19×10^{-7}
IPAH	26.7	-1.93, 63.7	0.070
PE	19.4	-9.26, 57.2	0.205
Male	7.11	-5.65, 21.6	0.288
Age	0.584	0.180, 0.990	0.005
Batch	7.33	-9.10, 26.7	0.403
Non-Caucasian	-14.5	-30.0, 4.42	0.124

Table 4.5 Multivariable linear regression model of VWF antigen levels

Reference groups are the same as described in Table 4.4. n=343 individuals included in the model.

There was a modest negative correlation between ADAMTS13 and VWF plasma levels in CTEPH ($\rho = -0.164$, $p=0.018$) but they were not correlated in healthy controls ($\rho = -0.0622$, $p=0.614$) (Figure 4.2). Furthermore, adjusting the multivariable linear regression model of ADAMTS13 antigen levels (Table 4.4) by VWF had minimal effect (on the β value), suggesting that the associations are not mediated by VWF antigen levels.

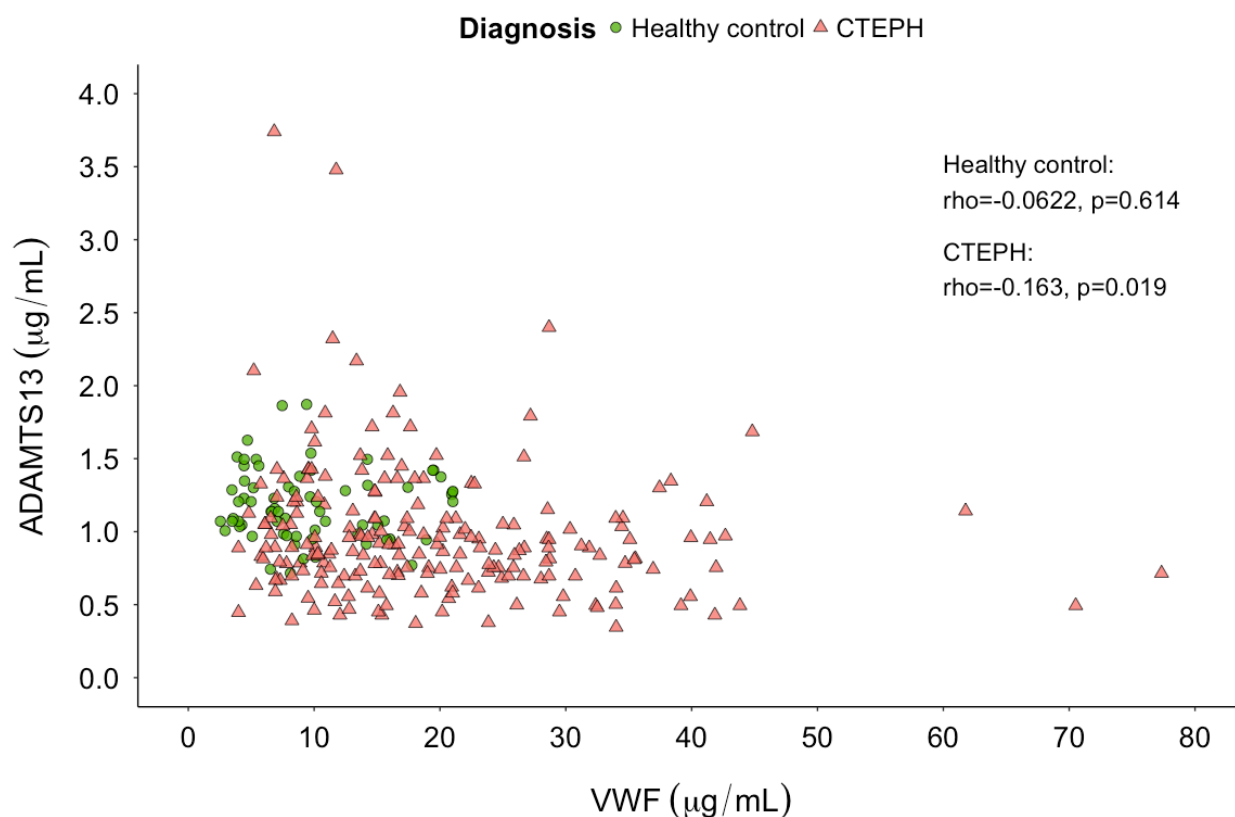


Figure 4.2. Correlation of ADAMTS13 with VWF antigen levels in CTEPH (n=208) and healthy controls (n=68)

Spearman's rank correlation in other diagnostic groups: CTED (rho = -0.161, $p = 0.354$), IPAH (rho = 0.329, $p = 0.076$), PE (rho = -0.0504, $p = 0.799$).

Converting ADAMTS13 antigen levels to a percentage of the median value of the healthy control group (set at 100%) allowed comparisons with thrombotic diseases in other studies. The majority of the CTEPH group (n=136, 65%) were in the lowest quartile (Q1<88% ADAMTS13) ([Table 4.6](#)).

The combination of low ADAMTS13 and high VWF antigen levels has a synergistic effect on the odds of CTEPH (Odds ratio (OR) = 14.5, $p<0.001$) compared with healthy controls ([Figure 4.3](#) and [Table 4.7](#)).

	CTEPH (n=208)	Healthy control (n=68)
Q1 (<88%)	136 (65)	16 (24)
Q2 (88-100%)	24 (12)	18 (26)
Q3 (100-114%)	12 (6)	17 (25)
Q4 (>114%)	36 (17)	17 (25)

Table 4.6 ADAMTS13 antigen level quartiles for CTEPH and healthy controls

ADAMTS13 levels were divided by the median of the healthy control group and expressed as a percentage. The CTEPH group was then divided into quartiles (Q1-Q4) of the ADAMTS13 distribution of the healthy control group.

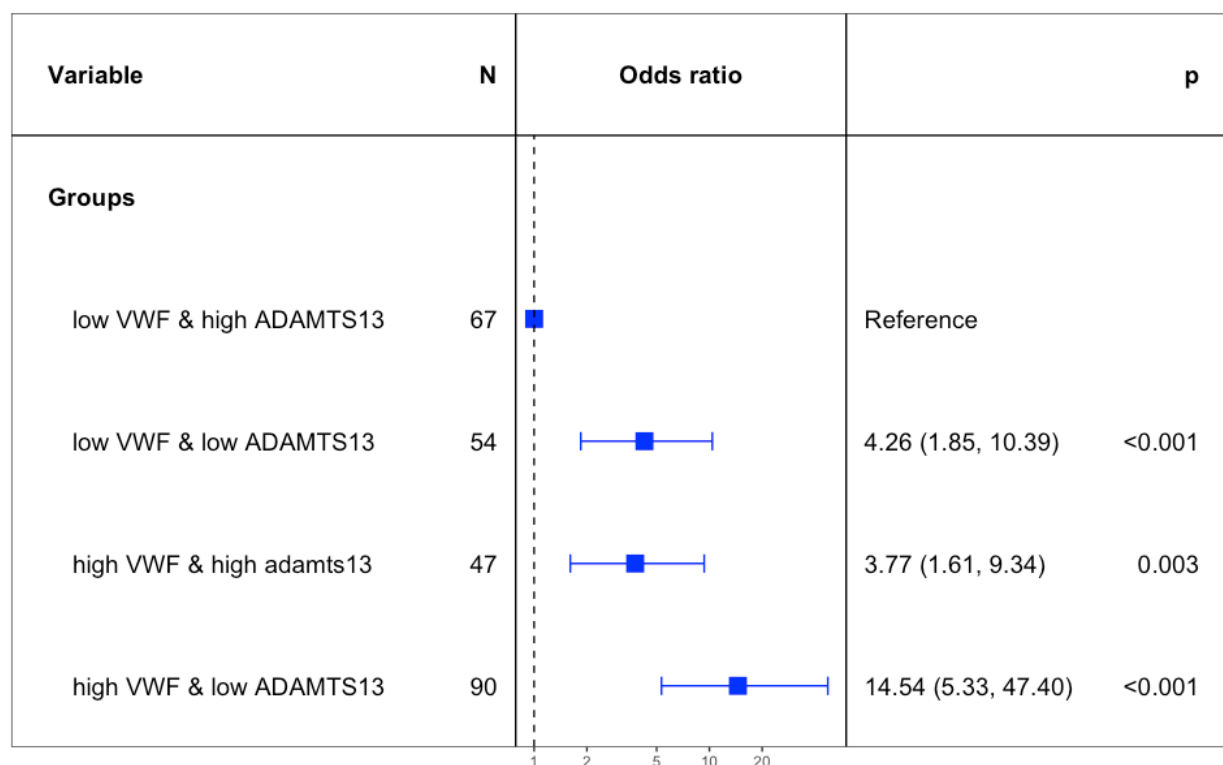


Figure 4.3 The odds ratios of CTEPH in relation to healthy controls for combined ADAMTS13 and VWF groups

The ADAMTS13 and VWF group ORs are adjusted for age, sex, ethnicity and batch in a logistic regression model. N represents the total for both the CTEPH and healthy control

groups. Threshold criteria and n (%) within the CTEPH and healthy control groups are shown in [Table 4.7](#). Forest plot generated with the R package `forestmodel`.(247)

Groups	Thresholds	Healthy Control	CTEPH
low VWF & high ADAMTS13	VWF: < 165% (Q1-Q3) ADAMTS13 > 88% (Q2-Q4)	38 (56)	33 (16)
low VWF & low ADAMTS13	VWF: < 165% (Q1-Q3) ADAMTS13: <= 88% (Q1)	13 (19)	45 (22)
high VWF & high ADAMTS13	VWF: >= 165% (Q4) ADAMTS13: > 88% (Q2-4)	12 (18)	39 (19)
high VWF & low ADAMTS13	VWF: >= 165% (Q4) ADAMTS13: <=88% (Q1)	5 (7)	91 (44)

Table 4.7 Summary table for combined ADAMTS13 and VWF groups

Threshold criteria are described in the table with the n (%) for CTEPH and healthy controls. ADAMTS13 and VWF antigen levels were converted to a percentage (the median of the healthy control group) and the quartile thresholds were then determined (healthy control group). The lowest quartile (Q1) was used to represent “low ADAMTS13” and the other quartiles (Q2-4) were used for “high ADAMTS13”. The highest quartile (Q4) was used to represent “high VWF” and the other quartiles (Q1-3) were used for “low VWF”. These cut points were used (as opposed to the median) due to the skewed distribution of ADAMTS13-VWF ([Table 4.6](#)) and to enable a comparison with other published thrombotic diseases that used a similar methodology.(212) The odds ratios for each group are shown in [Figure 4.3](#). The combined numbers in each group vary from [Figure 4.3](#), which used additional variables in a logistic regression model, some of which were missing.

4.2.2.3 Interaction effects

Interaction effects for the variables used in [Tables 4.4](#) and [4.5](#) were investigated. For ADAMTS13 antigen levels, there was a significant interaction between age and CTEPH ($p=0.007$) and additionally between age and sex (Age:Sex) ($p=0.019$) ([Table 4.8](#) and [Figure 4.4](#)). There was insufficient sample size to investigate 3-way interactions, including what was driving the interaction effect of age and sex. This suggests that the reduction in ADAMTS13 levels with increasing age is of more relevance within the CTEPH group. Consideration of the interaction terms is most relevant for the extreme values. For example, there is less difference between a 30-year-old Caucasian female with CTEPH and a 30-year-old Caucasian male healthy control (predicted ADAMTS13: 1.07 vs. 1.28 $\mu\text{g/mL}$, 16% reduction) than an 80-year-old Caucasian male with CTEPH and an 80-year-old Caucasian female healthy control (0.688 vs. 1.18 $\mu\text{g/mL}$, 42% reduction). There were no significant interaction effects for a separate model of VWF antigen levels using the variables in [Table 4.5](#).

	β	95% CI	p
(Intercept)	0.827	-0.194, 0.359	0.557
CTEPH	0.192	-0.135, 0.519	0.249
CTED	-0.325	-0.785, 0.136	0.167
IPAH	0.218	-0.236, 0.672	0.345
PE	-0.0728	-0.499, 0.353	0.737
Male	0.286	0.0305, 0.541	0.028
Age	0.00101	-0.00386, 0.00587	0.684
Batch	-0.0237	-0.111, 0.0638	0.594
Non-Caucasian	-0.0535	-0.16, 0.0527	0.322
CTEPH:Age	-0.00768	-0.0133, -0.00207	0.007
CTED:Age	0.000386	-0.00766, 0.00844	0.925
IPAH:Age	-0.00420	-0.012, 0.00358	0.289
PE:Age	-0.000410	-0.00825, 0.00743	0.918
Age:Sex	-0.00500	-0.00917, -0.000817	0.019

Table 4.8 Multivariable linear regression of ADAMTS13 plasma levels and interaction effects

The reference groups for diagnostic group, batch and ethnicity are the same as described in [Table 4.4](#). The interaction terms included in the model are those that were significant ($p<0.05$) and informative from the combination of variables in [Table 4.4](#). The beta coefficients should be interpreted with consideration of the interaction effects. For example, the predicted ADAMTS13 antigen level in an 80 year old male Caucasian with CTEPH from experimental batch1 would be: $\exp(0.0827 + 0.192 + (80 \times 0.00100) + (0.286) + (80 \times -0.00768) + (80 \times -0.00500)) = 0.688 \mu\text{g/mL}$. This is 34% lower than an 80 year old male Caucasian healthy control from experimental batch1: $\exp(0.827 + (80 \times 0.00100) + 0.286 + (80 \times -0.00500)) = 1.04 \mu\text{g/mL}$. $n=343$ included in the model.

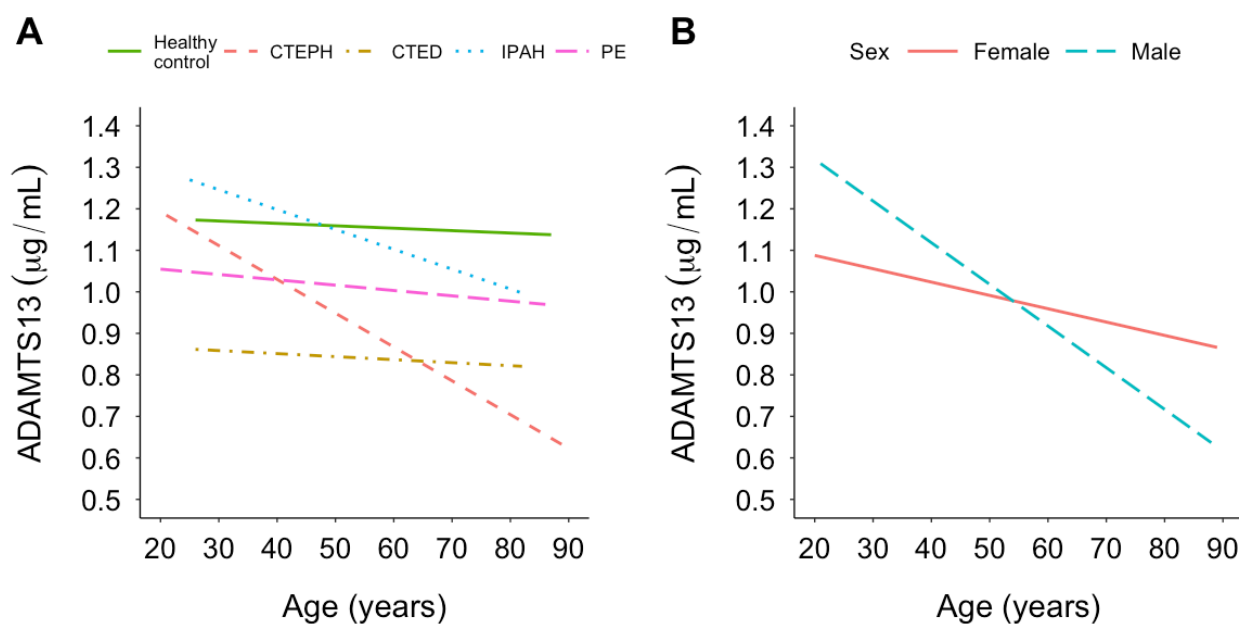


Figure 4.4 ADAMTS13 multivariable linear model interaction effects

Interaction effects for the model described in [Table 4.8](#). The predicted ADAMTS13 values are plotted with the interaction terms:

A. Age:Diagnostic group.

The ADAMTS13 difference between CTEPH patients and healthy controls is more pronounced for older patients than for younger patients. ADAMTS13 levels remain lower in CTEPH patients across all ages compared with healthy controls. The rate of ADAMTS13 reduction in the CTED group is similar to healthy controls and is also lower than healthy controls across all ages.

B. Sex:Age. n=343 individuals (from all groups) included in models

The rate of ADAMTS13 reduction with age is more pronounced for males than for females.

4.2.2.4 Replicates and ADAMTS13 batch adjustment

The median of the differences in ADAMTS13 protein levels between the replicates (n=24) of batch1 and batch2 (0.19 µg/mL; $p<0.001$) was used to adjust each batch2 ADAMTS13 value. This batch correction resulted in a larger CTEPH group size for subsequent genotype analyses. It was consistent with the ADAMTS13 antigen concentrations that have been previously reported with the same methodology.(212) As batch2 only contained CTEPH patients, no adjustment was applied to the other diagnostic groups. The validity of this approach was assessed with a multivariable linear model using the uncorrected ADAMTS13 values from batch2 together with values from batch1 (the dependent variable) and including the covariates batch, age, sex, ethnicity and diagnostic group (**Table 4.9**). This confirmed that the β coefficients and p -values were similar to the previous ADAMTS13 (corrected) multivariable linear model (**Table 4.4**) with the expected addition of a difference with batch ($\beta=+35.1\%$ (for batch2 with respect to batch1); $p<0.001$). Furthermore, the findings are maintained when limiting the analysis to the data from batch1 (**Figure 4.5**).

The median of differences in VWF protein levels between the replicates (n=12) of batch1 and batch2 was not significantly different (-0.98 µg/mL; $p=0.970$) and therefore a correction factor was not applied.

	β (%)	95% CI (%)	<i>p</i>
CTEPH	-23.4	-30.9, -15.1	5.91×10^{-7}
CTED	-25.9	-35.1, -15.4	1.18×10^{-5}
IPAH	-2.20	-14.7, 12.2	0.752
PE	-12.0	-24.0, 2.00	0.089
Batch	35.1	23.6, 47.7	1.20×10^{-10}
Age	-0.500	-0.700, -0.300	3.30×10^{-6}
Male	-1.10	-7.60, 5.90	0.756
Non-Caucasian	-5.60	-15.2, 5.10	0.293

Table 4.9 Multivariable linear regression model of uncorrected (batch2) ADAMTS13 antigen levels

Reference groups are the same as described in [Table 4.4](#). n=343 individuals included in the model.

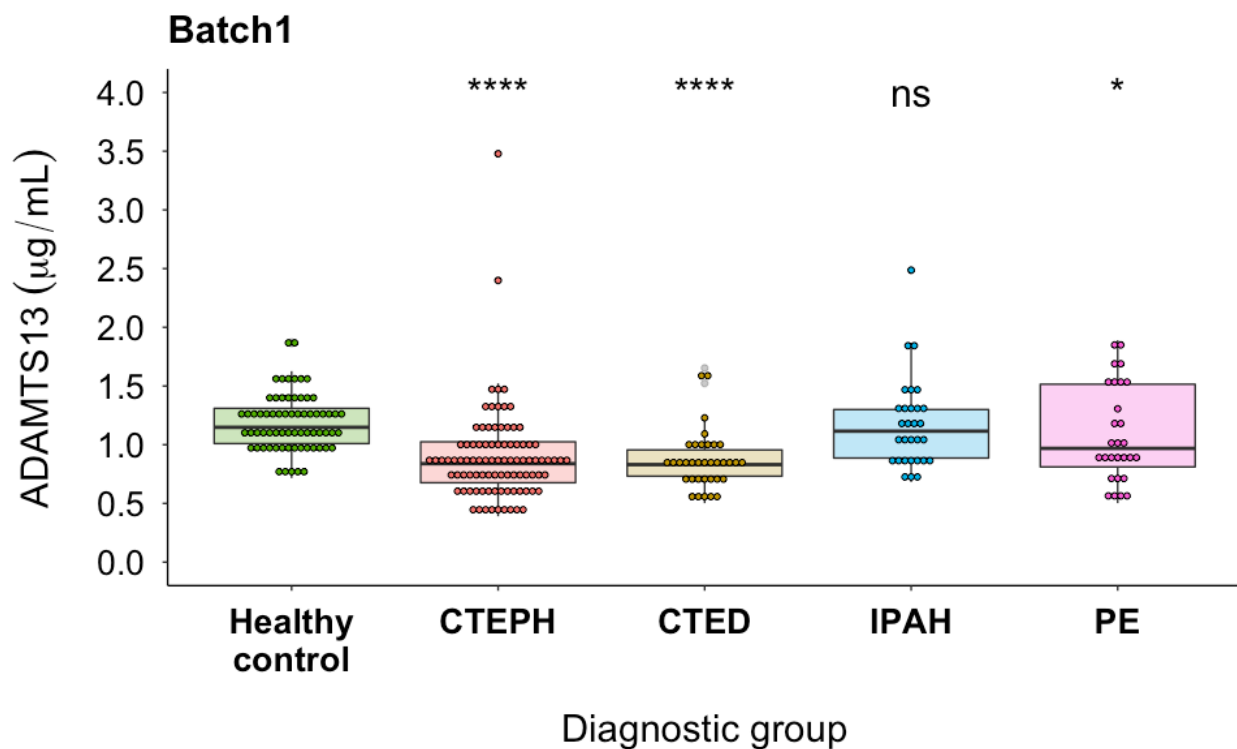


Figure 4.5 ADAMTS13 antigen levels by diagnostic groups for batch1

ADAMTS13 plasma levels for batch1 study participants *without* batch2 adjusted values for the CTEPH group (n=93). n: Healthy control=68, CTED=35, IPAH=28, PE=28.

4.2.2.5 ADAMTS13 and VWF: pre- and post-pulmonary endarterectomy

In 22 CTEPH patients matched samples were taken post-PEA, after a median of 343 days. There were no differences in ADAMTS13 (median of differences \pm IQR: -0.0328 ± 0.250 $\mu\text{g/mL}$, $p=0.777$) or VWF protein levels (-3.05 ± 10.7 $\mu\text{g/mL}$, $p=0.777$) following removal of proximal organised thrombus material by PEA ([Figure 4.6](#)).

ADAMTS13 and VWF levels did not change pre- and post-PEA and this also applied when limited to patients with normal post-operative haemodynamics (mPAP <25mmHg) (n=7, ADAMTS13: $p=0.742$, VWF: $p=0.195$).

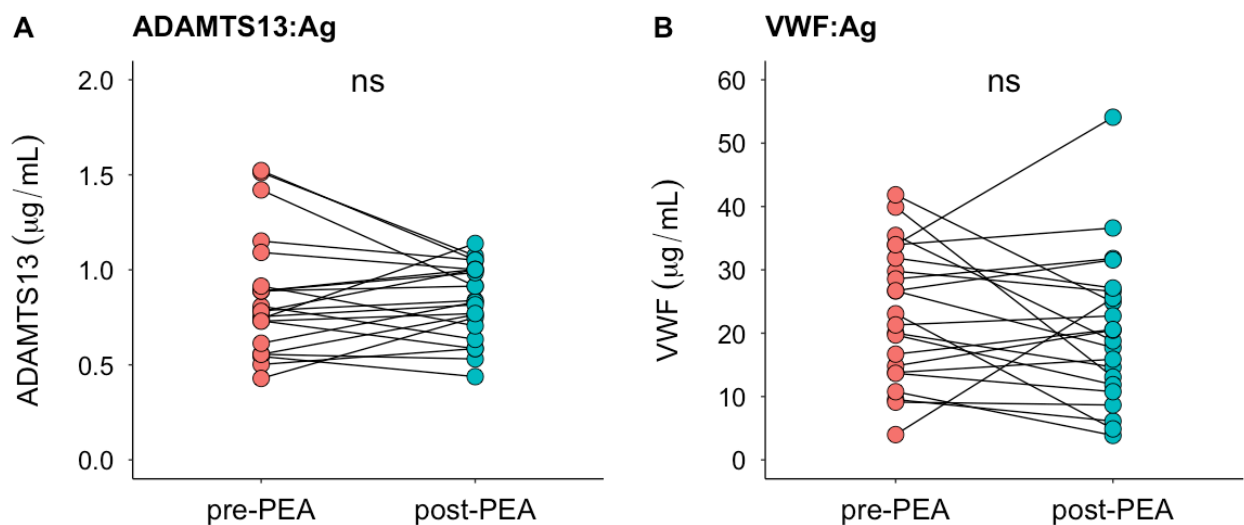


Figure 4.6 ADAMTS13 and VWF antigen levels pre- and post-pulmonary endarterectomy

A. ADAMTS13 and **B.** VWF antigen levels (n=22). Wilcoxon signed-rank test was used to calculate p -values.

4.2.3 ADAMTS13 activity, D-dimer and VWF multimers

ADAMTS13 activity and D-dimer concentrations were measured in a subset of patients with CTEPH (n=23) with the lowest ADAMTS13 protein concentrations (below the 1st Quartile of the CTEPH group) ($0.556 \pm 0.130 \mu\text{g/mL}$) and compared to a subset of healthy controls (n=14, ADAMTS13: $1.03 \pm 0.284 \mu\text{g/mL}$).

VWF multimeric size was assessed in a subset of CTEPH (n=21) with the highest VWF protein concentration (above the 3rd quartile) ($32.5 \pm 6.80 \mu\text{g/mL}$) and compared to the same subset of healthy controls (n=14, VWF: $9.97 \pm 4.99 \mu\text{g/mL}$).

4.2.3.1 ADAMTS13 activity

ADAMTS13 activity was reduced in CTEPH ($84 \pm 15\%$) compared with healthy controls ($107 \pm 14\%$; $p < 0.001$) ([Figure 4.7A](#)).

The ADAMTS13 FRET assay is influenced by ADAMTS13 antigen levels, with apparently low ADAMTS13 activity, if antigen levels are reduced. Therefore, the ADAMTS13 activity was adjusted for the ADAMTS13 antigen levels in each individual sample. Specific ADAMTS13 activity (Activity:antigen (Act:Ag) ratio) was increased in CTEPH (Act:Ag 1.57 ± 0.32) compared with healthy controls (1.05 ± 0.190 ; $p < 0.001$) ([Figure 4.7B](#)). Specific ADAMTS13 activity (Act:Ag) is not correlated with VWF:Ag in either CTEPH or healthy controls ([Figure 4.8A](#)).

4.2.3.2 D-dimers

Plasmin and thrombin are able to inactivate ADAMTS13 proteolytically *in vitro* and plasmin mediated ADAMTS13 cleavage has been observed in TTP.(283, 284) Furthermore, abnormalities in the fibrinolysis pathway have been implicated in CTEPH.(39) Therefore, fibrinogen degradation products measured by D-dimer were used as a potential surrogate marker of plasmin and thrombin activity. D-dimer was increased in CTEPH ($1.24 \pm 1.25 \mu\text{g/mL}$) compared to healthy controls ($0.538 \pm 0.344 \mu\text{g/mL}$; $p = 0.030$) ([Figure 4.7C](#)). Specific ADAMTS13 activity was not correlated with D-

dimer in the CTEPH ($\rho=0.0938$, $p=0.761$) or healthy control groups ($\rho=-0.220$, $p=0.313$) (**Figure 4.7D**).

4.2.3.3 VWF multimeric size

It was hypothesised that a decrease in ADAMTS13 antigen levels would result in reduced VWF cleavage and an increase in high multimeric VWF as occurs in TTP.(285) There was no difference in VWF multimeric size between CTEPH (VWF CBA:Ag ratio, 0.659 ± 0.537) and healthy controls (0.866 ± 0.494 ; $p=0.160$) (**Figure 4.7E**). VWF:CBA was not correlated with ADAMTS13:Ag in CTEPH or healthy controls (**Figure 4.8B**). VWF:CBA was correlated with VWF:Ag in healthy controls but not in CTEPH (**Figure 4.8C**).

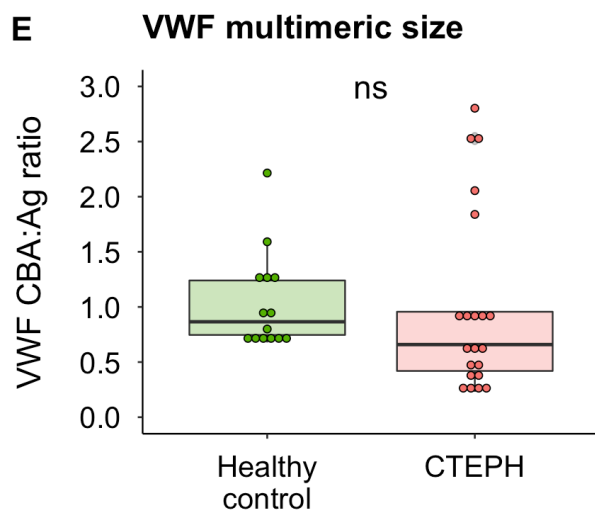
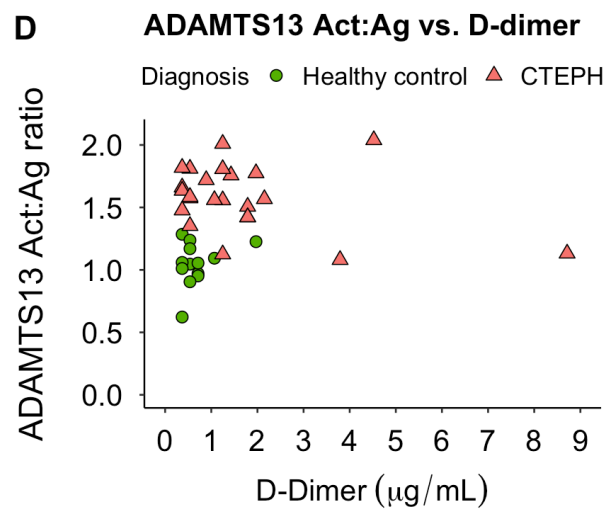
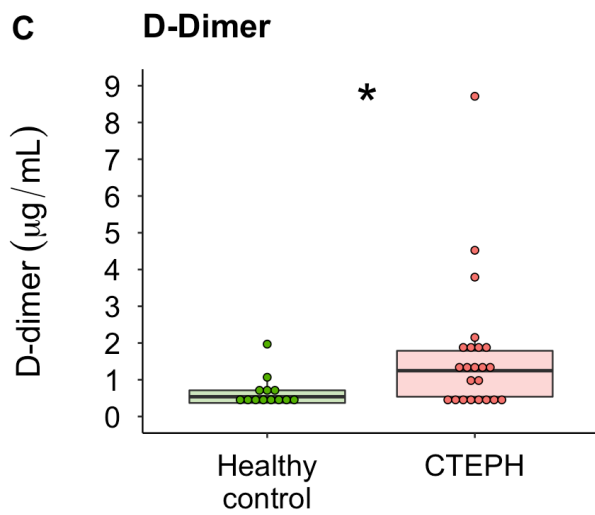
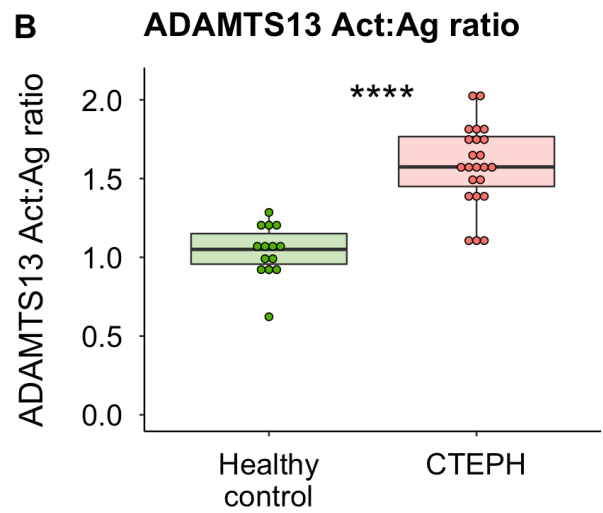
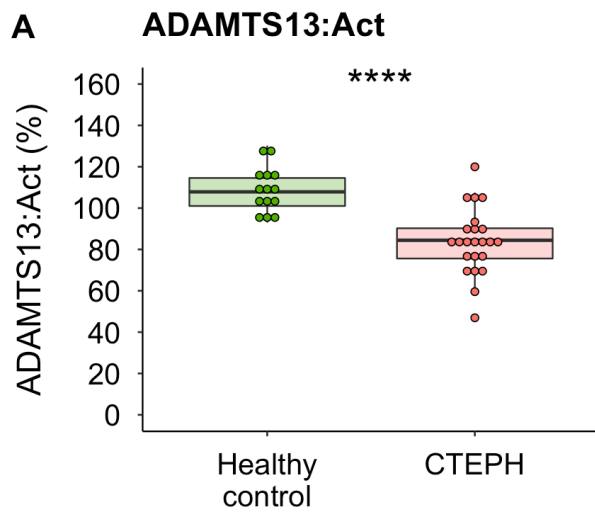


Figure 4.7 ADAMTS13 activity, D-dimer and VWF multimeric size in CTEPH and healthy controls

A subset of CTEPH patients (n=23) with the lowest ADAMTS13 antigen levels (below the first quartile of the CTEPH group) and healthy controls (n=14) were used for **A-D**. VWF multimeric size was measured in CTEPH (n=21) samples with the highest VWF antigen concentrations (above the third quartile of the CTEPH group) using the same healthy control subset and displayed in **E**. The Mann-Whitney *U* test was used to calculate group differences (**A, B, C, E**) and correlation was assessed with Spearman's rank correlation coefficients (**D**). **A.** ADAMTS13 activity (%) **B.** Specific ADAMTS13 activity (Act:Ag ratio). **C.** D-dimer antigen levels. **D.** Specific ADAMTS13 activity and D-dimer antigen correlation. Healthy control correlation: $\rho=0.0938$, $p=0.761$; CTEPH correlation: $\rho=-0.220$, $p=0.313$. **E.** VWF multimeric size (VWF Collagen binding assay:Antigen ratio).

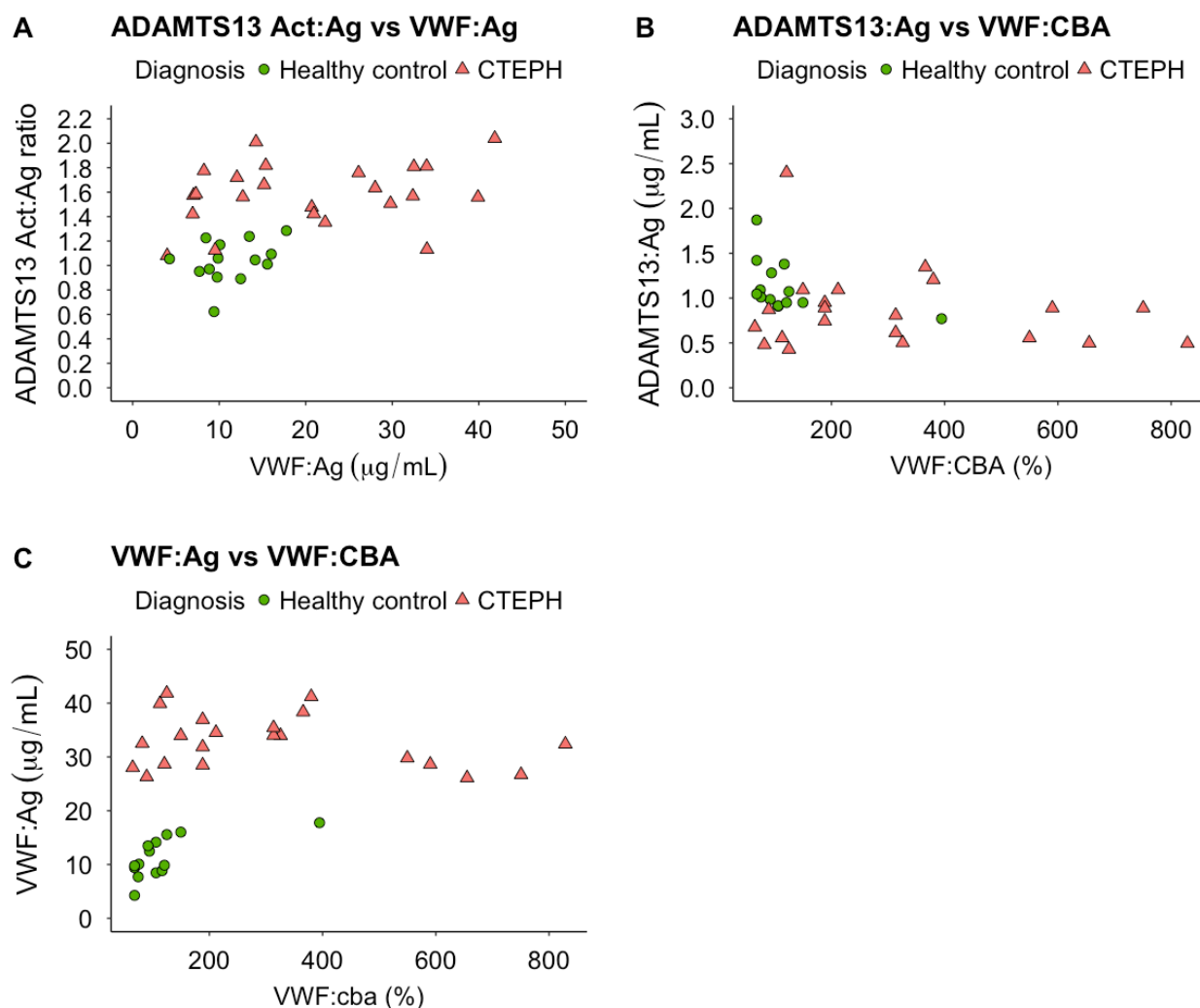


Figure 4.8 ADAMTS13 activity, D-dimer and VWF multimeric size correlation in CTEPH and healthy controls

Correlation was assessed with Spearman's rank correlation coefficients for both the healthy control (green circles) and the CTEPH (red triangles) groups. Additional analysis from the data shown in [Figure 4.7](#) (See [Figure 4.7](#) for group details).

A. ADAMTS13 Act:Ag ratio vs. VWF:Ag. Healthy control correlation: $\rho=0.36$, $p=0.210$; CTEPH correlation: $\rho=0.24$, $p=0.270$.

B. ADAMTS13:Ag ratio vs. VWF:CBA. Healthy control correlation: $\rho=-0.61$, $p=0.022$; CTEPH correlation: $\rho=-0.02$, $p=0.930$.

C. VWF:Ag vs. VWF:CBA. Healthy control correlation: $\rho=0.670$, $p=0.008$; CTEPH correlation: $\rho=-0.088$, $p=0.700$.

4.2.4 Clinical phenotype associations with ADAMTS13 and VWF

The association between clinical phenotypes and ADAMTS13 or VWF antigen levels was assessed in all patients with CTEPH (n=208). ADAMTS13 and VWF did not significantly correlate with pulmonary vascular resistance, 6-minute walk distance or N-terminal pro b-type natriuretic peptide (NT-proBNP), which are markers of disease severity ([Figure 4.9](#)). Since inflammation has been associated with both CTEPH and abnormalities in the ADAMTS13-VWF axis, correlation between the ADAMTS13-VWF axis and inflammatory markers was investigated.(46, 286) There were no correlations with blood markers of inflammation (CRP, WCC, neutrophil and lymphocyte %) ([Figure 4.10](#)), including when confining the analysis to samples that were taken on the same day as ADAMTS13 and VWF sampling (n=81, for WCC, neutrophil and lymphocyte %; n= 77 for CRP).

As proximal operable CTEPH has different risk associations to distal inoperable CTEPH and thus potentially different pathophysiological mechanisms, the disease sub-types were investigated.(78) There was no difference in ADAMTS13 ($p=0.070$) or VWF ($p=0.253$) between the different sub-diagnostic categories of CTEPH ([Figure 4.11A](#) and [4.11B](#)). Furthermore, there was no difference in ADAMTS13 ($p=0.366$) or VWF ($p=0.078$) in those with and without post-operative residual pulmonary hypertension (mPAP \geq 25mmHg) (n=83, 63%), which is a potential marker of distal vasculopathy ([Figure 4.11C](#) and [4.11D](#)).(15)

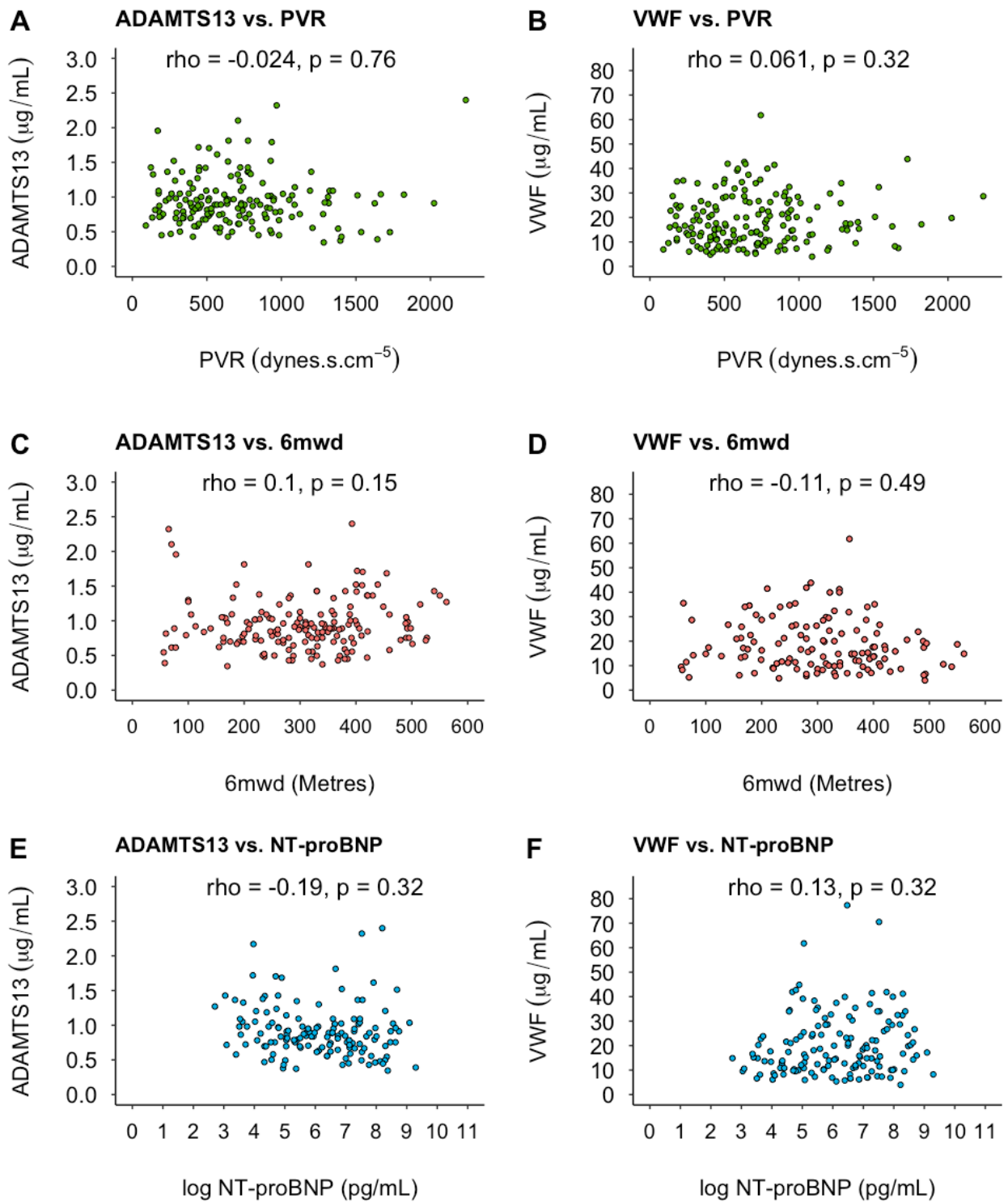


Figure 4.9 Correlation of ADAMTS13 and VWF antigen levels with markers of disease severity in CTEPH at baseline

Correlation was assessed by Spearman's rank test. Baseline was defined as the value closest to diagnosis. *P*-values adjusted for the number of statistical tests performed using

FDR correction. 6mwd (6-minute walk distance), NT-proBNP (N-terminal pro b-type natriuretic peptide), PVR (pulmonary vascular resistance). NT-proBNP log-transformed to improve visualisation. Numbers per group: PVR = 169, 6mwd = 165, NT-proBNP = 144.

CTEPH is a potential severe consequence of acute PE, however there are a spectrum of changes following PE (post-PE syndrome) that may have differing pathobiology.(9) ADAMTS13 and VWF antigen levels were evaluated in groups with varying degrees of post-PE perfusion defects on available VQ scans (n=20). There was no difference in ADAMTS13 ($p=0.812$) or VWF ($p=0.678$) levels in those that had residual perfusion defects post-PE (n=12) compared with those with no perfusion defects (n=8) (**Figure 4.12A** and **4.12B**). Furthermore, there was no difference in ADAMTS13 ($p=0.938$) or VWF ($p=0.427$) levels when the PE group was stratified into provoked PE (n=8) and idiopathic PE (n=12) (**Figure 4.12C** and **4.12D**).

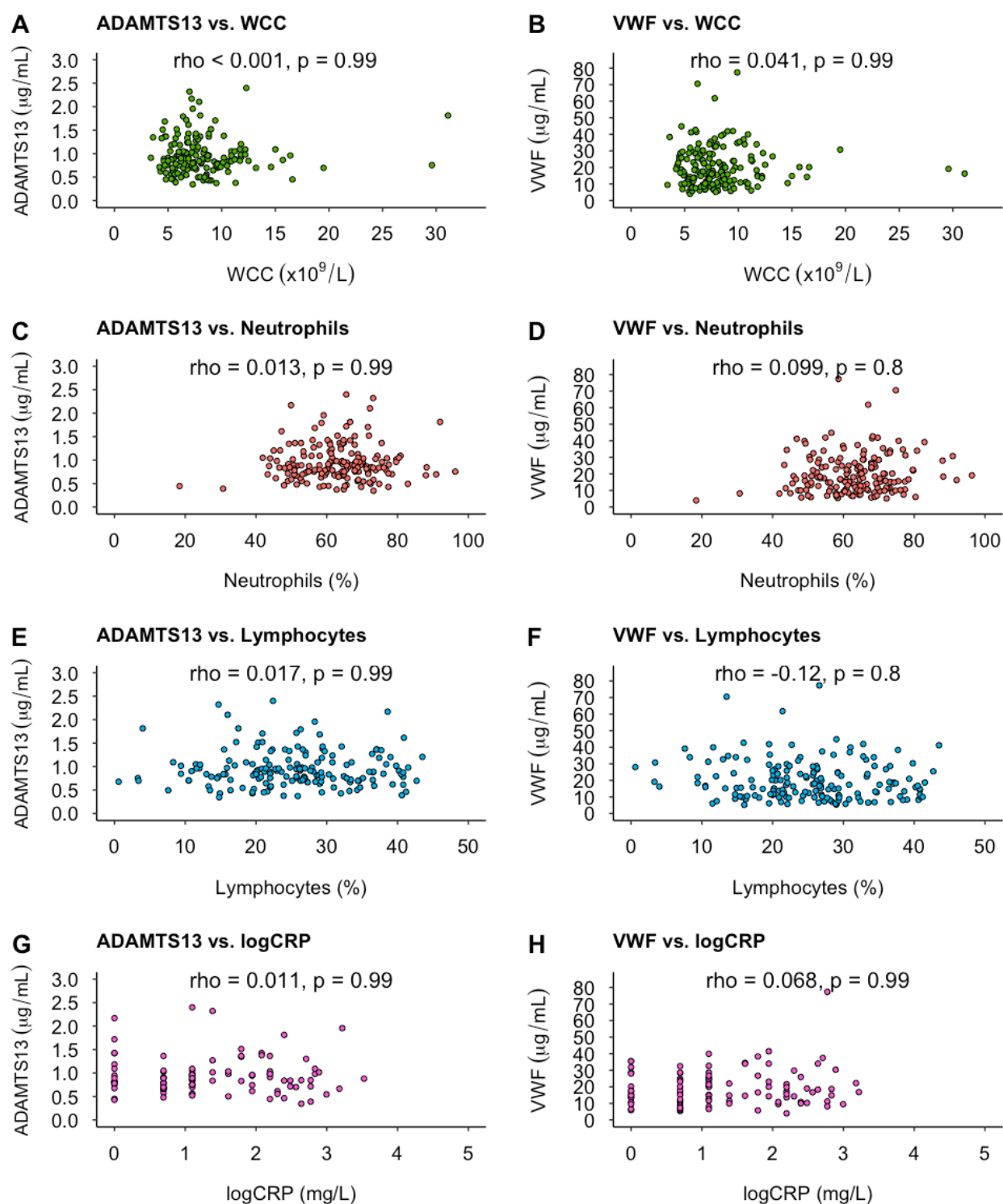


Figure 4.10 Correlation of ADAMTS13 and VWF antigen levels with blood markers of inflammation at baseline

Correlation was assessed by Spearman's rank test. *P*-values adjusted for the number of statistical tests performed using FDR correction. CRP log-transformed to improve visualisation. CRP (C-reactive protein), WCC (white cell count). Numbers in each group: WCC = 169, Neutrophils = 169, Lymphocytes = 168, CRP = 95.

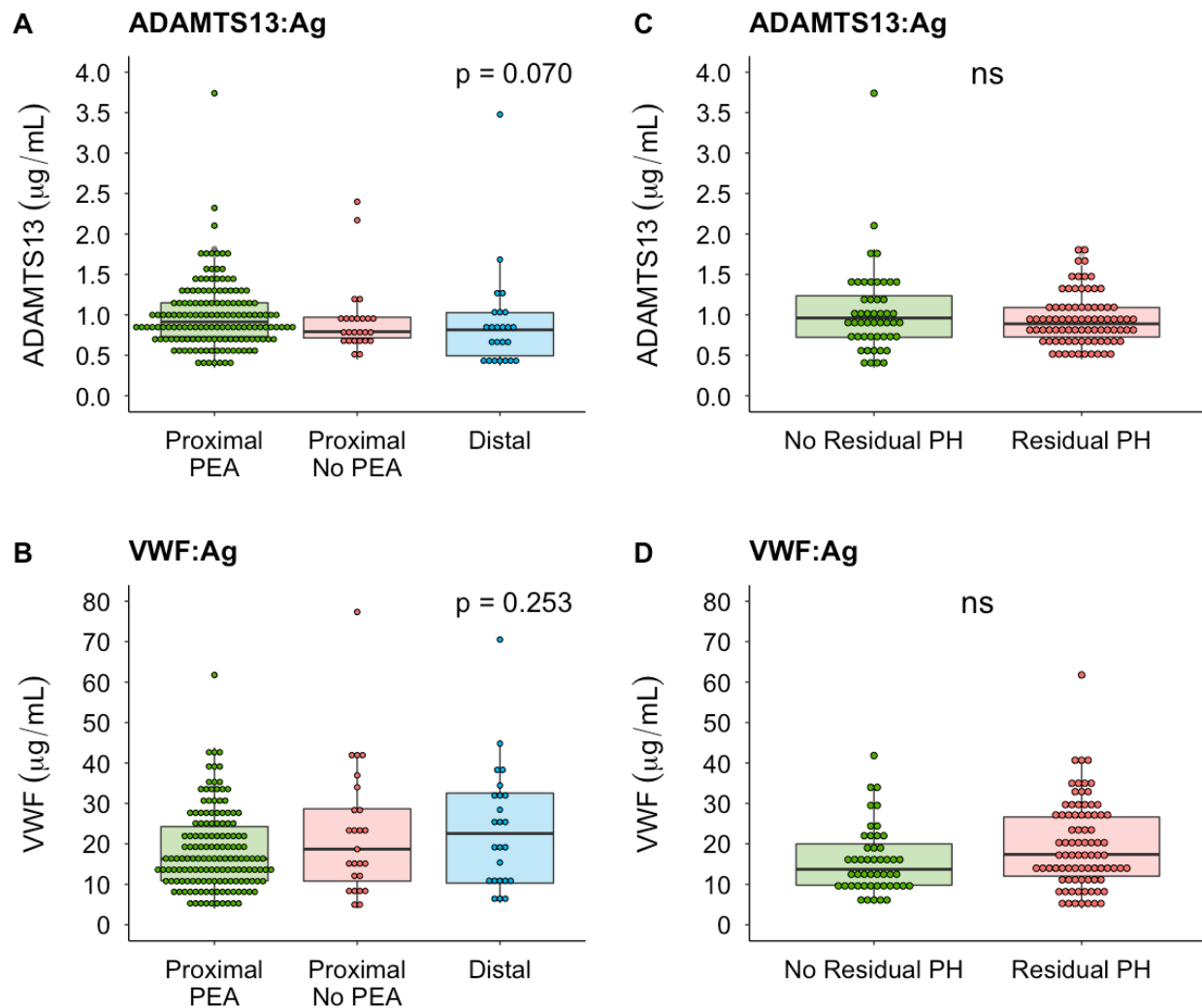


Figure 4.11 ADAMTS13 and VWF antigen levels in CTEPH sub-diagnostic and post-PEA residual pulmonary hypertension groups

A, B. ADAMTS13 and VWF antigen levels in CTEPH diagnostic sub-groups. Numbers in each group: proximal PEA = 150, proximal no PEA = 25, distal (surgically inaccessible) = 24, insufficient clinical data in 9 patients to classify them.

C, D. ADAMTS13 and VWF antigen levels in CTEPH post-PEA residual pulmonary hypertension (mPAP ≥ 25 mmHg) groups. Numbers in each group: no residual PH = 49, residual PH = 83, insufficient clinical data in 18 patients to classify them. The group differences were assessed using the Kruskal-Wallis test (**A, B**) and the Mann-Whitney *U* test (**C, D**).

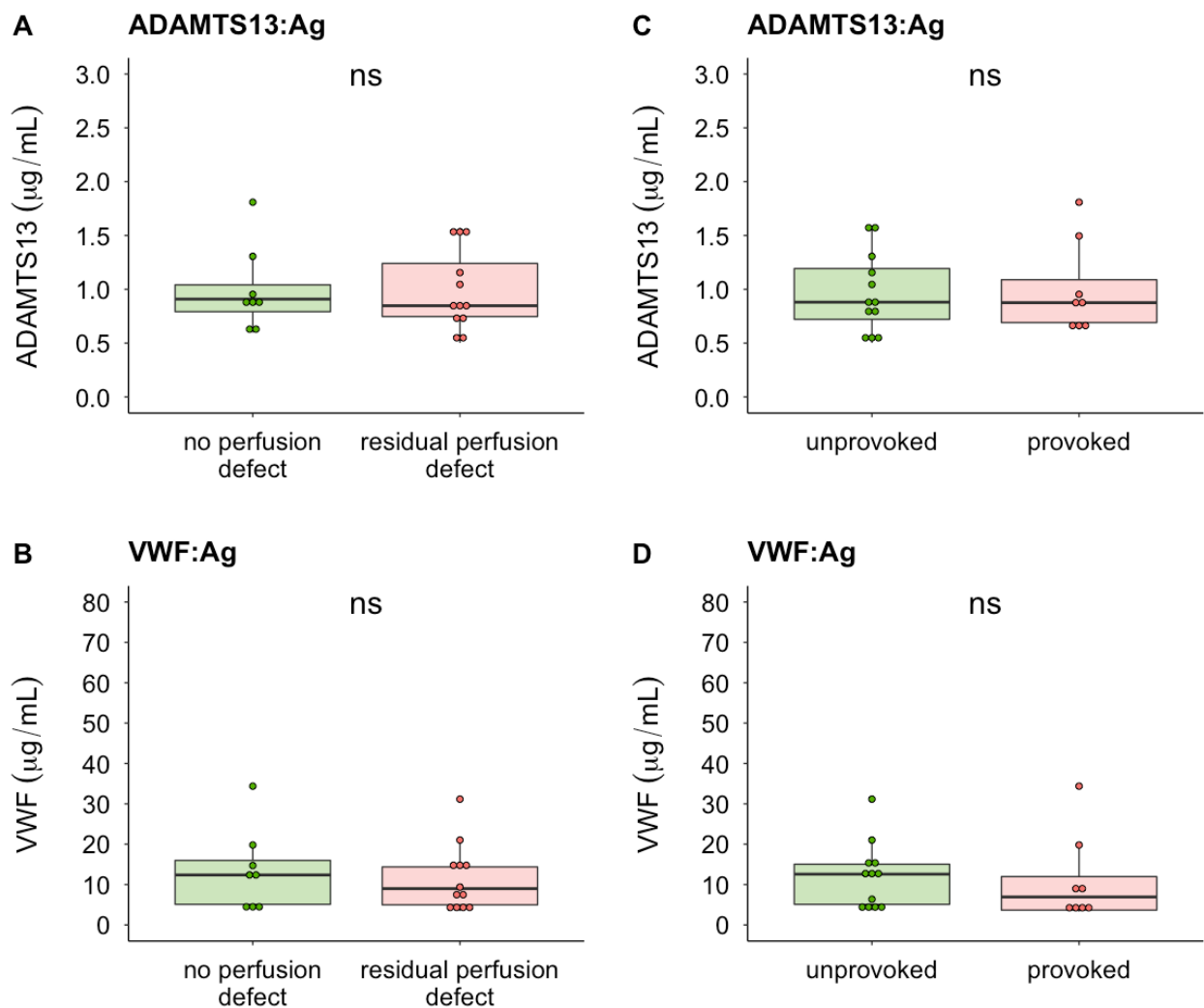


Figure 4.12 ADAMTS13 and VWF antigen levels in PE stratified by residual perfusion defects and provoked PE

The group differences were assessed using the Mann-Whitney *U* test. Numbers in each group: no perfusion defect on VQ scan = 8, residual perfusion defect on VQ scan = 12, unprovoked (no VTE risk factors) = 12, provoked (VTE risk factors) = 8. Of those with residual perfusion defects, the majority were minor (n=10).

4.2.5 ADAMTS13-VWF and genotype analyses

Imputed genotype dosages were available from the CTEPH GWAS described in [Section 2.1.6](#). All individuals were genotyped on commercially available Illumina assays and imputed to the Haplotype Reference Consortium Build 1.1.

208 CTEPH patients with ADAMTS13 / VWF antigen levels and 28 patients with CTED were also included in the CTEPH GWAS (CTED patients were not included in the final GWAS analysis and were removed due to incorrect phenotype, see [Section 3.2.2.2](#)). Genotypes were available for 207 (187 CTEPH; 23 CTED) after GWAS quality control exclusions. These patients were included in the genetic *ABO* group and protein quantitative trait loci (pQTL) analyses. Matched genotypes and ADAMTS13 / VWF antigen levels were not available for the healthy control, IPAH or PE groups.

4.2.5.1 Genetic *ABO* groups and ADAMTS13-VWF

Reconstructing genetic *ABO* groups allowed us to explore more complex associations within the *ABO* subgroups. Whilst the A1 and A2 groups would be classified as non-O on serological testing, they have been associated with different effects on VWF levels and VTE risk.⁽²⁵³⁾ There was no difference in ADAMTS13 antigen levels when stratified by simple genetic *ABO* groups (O, A, B, AB) ([Figure 4.13A](#)) ($p=0.443$) or more comprehensive genetic *ABO* groups ([Figure 4.14A](#)) ($p=0.616$).

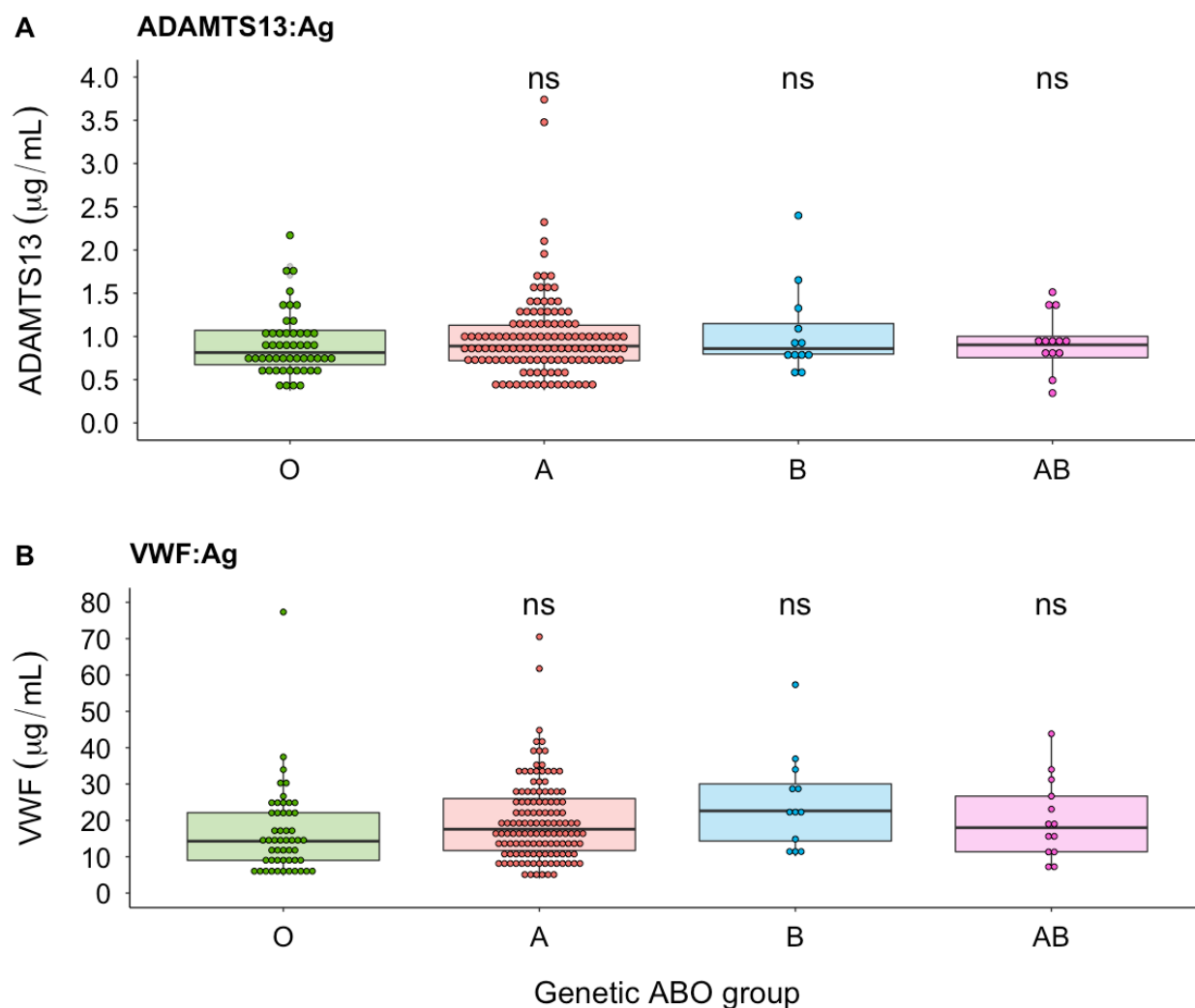


Figure 4.13 ADAMTS13 and VWF antigen levels by ABO genetic groups

CTEPH (n=182) and CTED (n=22) patients with genotypes and protein levels available (in n=3 a genetic ABO group could not be inferred) were included. Dunn's test with FDR adjustment was used to calculate *p*-values. Numbers in each group: O = 51, A = 128, B = 12, AB = 13.

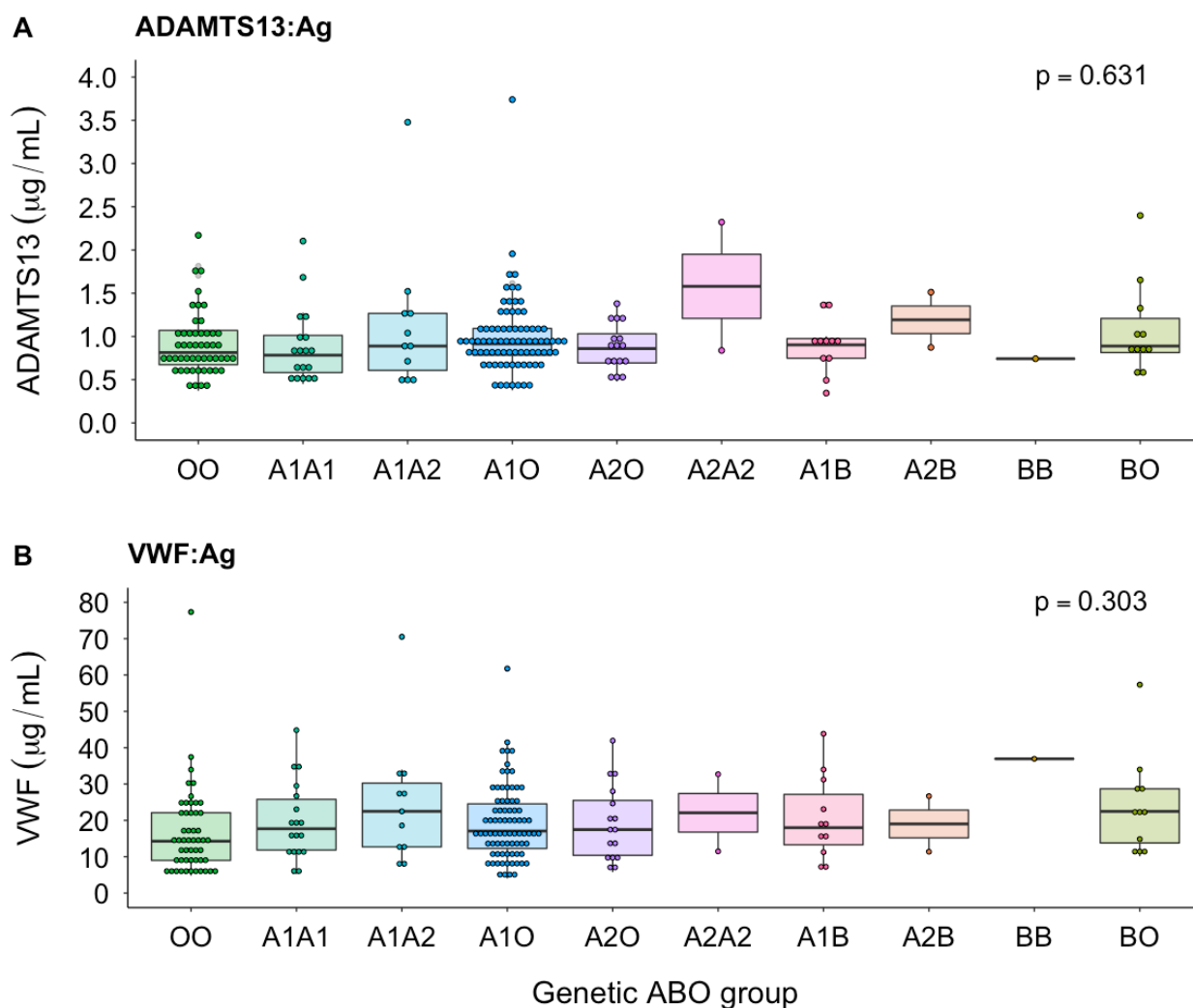


Figure 4.14 ADAMTS13 and VWF antigen levels by comprehensive ABO genetic groups

The group differences were assessed using the Kruskal-Wallis test. Numbers in each group: OO = 51, A1A1 = 18, A1A2 = 11, A1B = 11, A1O = 81, A2A2 = 2, A2B = 2, A2O = 16, BB = 1, BO = 11 (in $n=3$ a genetic ABO group could not be inferred).

VWF levels did not vary by ABO groups ([Figure 4.13B](#) and [Figure 4.14B](#)) however, when accounting for covariates ([Table 4.10](#)), ABO group B had a higher VWF level ($\beta=+51.3\%$, $p=0.025$) compared to group O. ABO group A also had a higher VWF level, although this was not statistically significant ($\beta=+19.8\%$, $p=0.073$). Patients with ABO group O had the lowest VWF levels within the CTEPH group ($14.5 \pm 13.0 \mu\text{g/mL}$), however this was still

significantly higher than healthy controls ($8.45 \pm 8.77 \mu\text{g/mL}$, $p < 0.001$) ([Figure 4.1](#)). This implies that the increase in VWF observed in the CTEPH group is not only driven by ABO.

	β (%)	95% CI (%)	p
ABO group - O	Reference		
ABO group - A	19.8	-1.75, 46.1	0.074
ABO group - B	51.3	5.30, 117	0.025
ABO group - AB	4.41	-26.9, 49.1	0.811
CTEPH	Reference		
CTED	7.43	-18.5, 41.6	0.609
Male	4.40	-11.9, 23.7	0.617
Age	0.921	0.341, 1.50	0.002
Batch	11.8	-5.88, 32.8	0.203
Non-Caucasian	-1.71	-51.7, 100	0.962

Table 4.10 Multivariable linear regression model of VWF antigen levels and genetic ABO groups in CTEPH

The reference ABO group is O and the reference diagnostic group is CTEPH. Otherwise, reference groups are the same as described in [Table 4.4](#). n=204 included in the model.

There was no difference in ADAMTS13 antigen levels between ABO groups, when accounting for covariates with multivariable linear regression.

4.2.5.2 Protein quantitative trait loci for ADAMTS13

There were 5 SNPs in the *ADAMTS13* \pm 40kb region that were significantly associated with ADAMTS13 protein in a linear regression model ([Table 4.11](#)). The most significant SNP (rs3739893, risk allele C, information score=0.960, β = -37.1%, $p=3.78 \times 10^{-06}$) is a 5' untranslated region (UTR) variant in the *C9orf96* gene, which is ~8kb 5' of the *ADAMTS13* gene. In a model adjusted for age, sex and batch, the lead SNP (rs3739893) explained 7.7% of the variance in ADAMTS13 levels within the CTEPH group ([Table](#)

4.12). However, as only 10 CTEPH patients had the rs3739893 effect allele, this accounts for a small proportion of the ADAMTS13 antigen level reduction observed in CTEPH. In the whole CTEPH GWAS cohort, the effect allele frequency for rs3739893 in CTEPH cases (0.0128) and healthy controls (0.0158) was not significantly different, which suggests that it is not associated with CTEPH disease risk. The effect allele frequency of the study healthy controls was similar to a European (non-Finnish) reference population (0.0160) (<http://gnomad.broadinstitute.org/>, accessed Feb 2018).

The most variance in ADAMTS13 antigen levels was attributable to age (16%) (Table 4.12), which is higher than reported in healthy cohorts.(231) The 4 other significant SNPs were highly correlated with the lead SNP ($R^2=0.91-1.00$, $p<0.001$). Additional analysis correcting for the first 5 ancestry informative principal components and VWF antigen levels did not alter the results. Furthermore, the results were unchanged when the analysis was confined to the CTEPH group.

rsID	CHR	Position	β (%)	95% CI (%)	p
rs3739893	9	136243324	-37.1	-48.1, -23.8	3.78×10^{-6}
rs28407036	9	136252654	-39.0	-51.3, -23.5	2.42×10^{-5}
rs8181039	9	136253927	-37.4	-49.5, -22.5	2.42×10^{-5}
rs78883179	9	136241818	-41.0	-54.1, -24.1	5.05×10^{-5}
rs77533110	9	136286789	-43.0	-56.4, -25.5	5.20×10^{-5}

Table 4.11 Protein quantitative trait loci for ADAMTS13 antigen levels in CTEPH

Associations were assessed using multivariable linear regression and the SNPs included were those in the *ADAMTS13* gene \pm 40 Kilobases ($n=396$). The model was adjusted for age, sex and batch. A Bonferroni p -value threshold $<1.26 \times 10^{-4}$ ($0.05/396$ variants) was used to denote statistical significance. rsID (reference SNP identification), CHR (chromosome), position (base position). GRCh37 was used for the genomic positions of the SNPs. $n=207$ individuals included in the model.

	β (%)	95% CI (%)	<i>p</i>	Variance (%)
rs3739893	-37.1	-48.1, -23.8	3.78×10^{-6}	7.70
Age	-0.935	-1.21, -0.660	2.23×10^{-10}	16.3
Male	-4.55	-11.7, 3.34	0.253	0.651
Batch	-2.40	-9.67, 5.47	0.537	0.0748

Table 4.12 Multivariable linear regression with the percentage of variance of ADAMTS13 antigen levels explained by SNPs and other characteristics

Reference groups are the same as described in [Table 4.4](#). Partitioning of the variance explained by each variable within the models was performed by averaging over orders using the R package `relaimpo`. (233) n=207 individuals included in the model.

4.2.6 Immunohistochemistry

Immunohistochemistry was assessed to determine if ADAMTS13 is expressed in vascular endothelial cells, which is a major site of disease pathogenesis. ADAMTS13 was expressed in the vascular endothelium of CTEPH and control lungs with no clear differences ([Figure 4.15](#)). Furthermore, ADAMTS13 was expressed in the endothelial neovessels of chronic thromboembolic material.

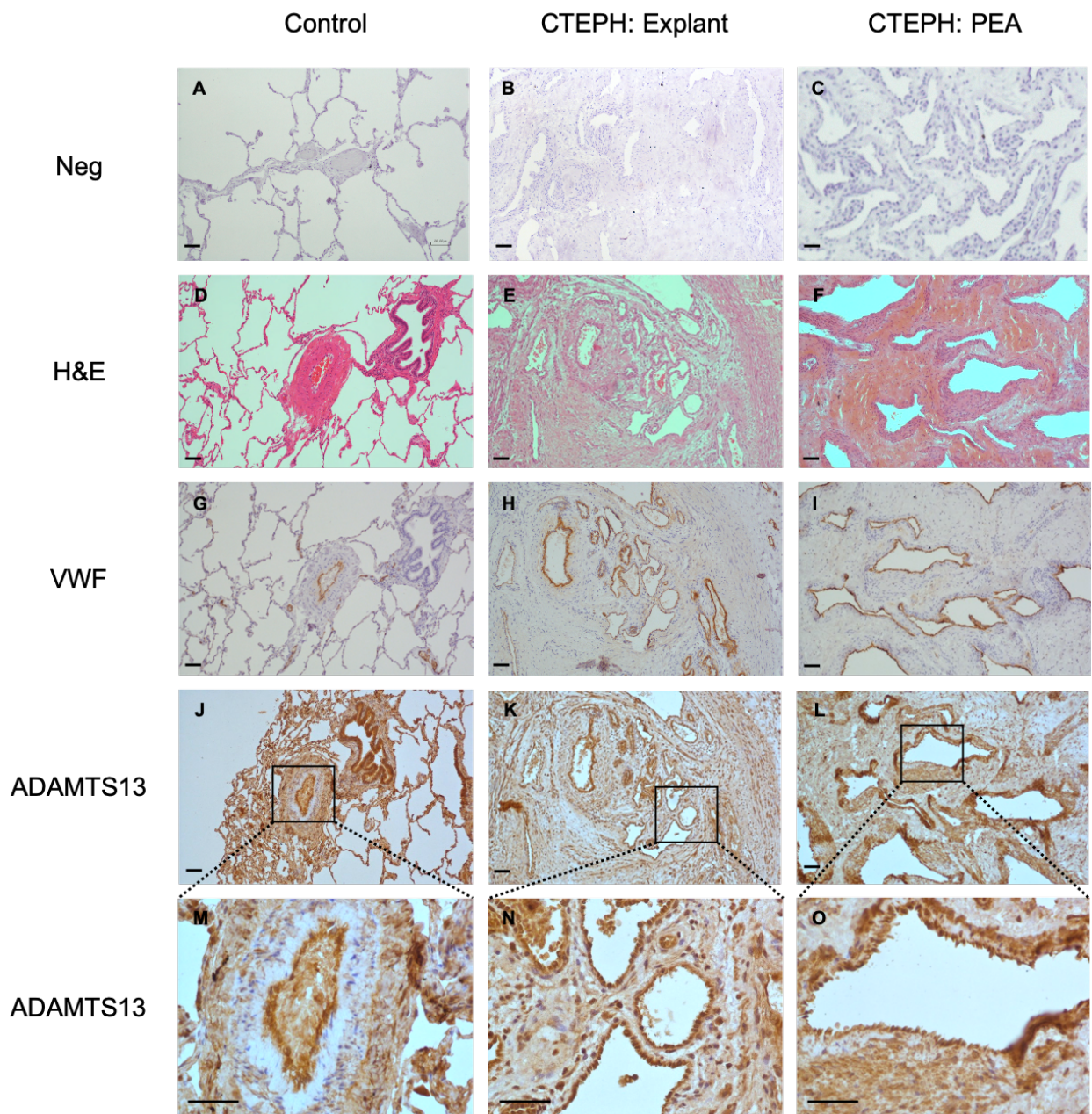


Figure 4.15 Expression of ADAMTS13 in lung tissue evaluated by immunohistochemistry

Serialised sections of lung tissue from controls, explanted lungs from patients with chronic thromboembolic pulmonary hypertension (CTEPH) and chronic thromboembolic material removed during pulmonary endarterectomy (PEA). Scale bars = 50 μ m. Negative (Neg) control (**A-C**) H&E staining (**D-F**); VWF (**G-I**); ADAMTS13 (**J-L**); ADAMTS13 high power (**M-O**)

field views (**M-O**). The negative control images are from the same tissue but a different location (as images from the same location were unavailable) and are representative.

4.3 Discussion

4.3.1 Overview

This is the first study demonstrating a marked reduction in levels of ADAMTS13 in CTEPH. This is independent of pulmonary hypertension, disease severity or systemic inflammation. VWF was confirmed to be increased in CTEPH and this implicates dysregulation of the ADAMTS13-VWF axis in CTEPH pathobiology.

4.3.2 ADAMTS13-VWF plasma levels and other diseases

The ADAMTS13-VWF axis is dysregulated in other thrombotic diseases including coronary artery disease and ischaemic stroke.(191, 193) However, the magnitude of ADAMTS13 reduction and VWF increase in CTEPH is greater than observed in studies of ischaemic stroke using the same methodology.(212) Furthermore, levels are lower in CTEPH than CAD when considering the proportion of patients in the lowest ADAMTS13 quartile (65% versus 28% respectively), although direct comparison is precluded by differing thresholds.(193) Additionally, the combination of decreased ADAMTS13 and increased VWF has a synergistic effect on the odds of CTEPH that is greater than observed in CAD or ischaemic stroke.(212) The more pronounced ADAMTS13-VWF dysregulation in CTEPH may reflect the larger surface area of the vascular endothelium involved or alternatively that ADAMTS13-VWF dysregulation is more important in CTEPH pathobiology. Although ADAMTS13 is predominately produced by the liver, the contribution to plasma levels from vascular endothelial cells could be substantial given the large surface area of the lung vasculature.(177)

Following PEA and removal of proximal thromboembolic material, the ADAMTS13-VWF axis remains dysregulated despite normalisation of haemodynamic parameters. Additionally, there is an equal perturbation of the axis in CTED, and no correlation with CTEPH disease severity, confirming the changes are not due to the presence of pulmonary hypertension or organised thrombus *per se*. Interestingly, there was no abnormality in ADAMTS13 levels in IPAH despite this group having a higher pulmonary vascular resistance, implying that distal pulmonary artery endothelial

dysfunction and small vessel vasculopathy are not responsible.(287) Taken together, these observations demonstrate the dysregulation of the ADAMTS13-VWF axis in CTEPH pathogenesis.

4.3.3 ADAMTS13-VWF: dysregulation mechanism and role in CTEPH pathobiology

It has been hypothesised that low ADAMTS13 is driven by activation of fibrinolytic pathways and an increase in thrombin and/or plasmin, which have the potential to proteolytically inactivate ADAMTS13.(283) Whilst D-dimer was raised in CTEPH there was no correlation with ADAMTS13 implying this was not the mechanism by which ADAMTS13 was reduced. High multimeric forms of VWF appear not to be increased in CTEPH. This is surprising, as increased high multimeric VWF occurs when ADAMTS13 is reduced in TTP and has been suggested to occur in ischaemic stroke and CAD.(212, 285) VWF multimeric size measured systemically may not reflect the local disease microenvironment in the pulmonary vascular endothelium. Additionally, localised flow conditions that may be altered in CTEPH are important in VWF structure, cleavage by ADAMTS13 and thrombus resolution.(178) Shear stress is required to unfold VWF and expose its A2 binding domain to ADAMTS13 however, pulmonary arterial shear stress is decreased in pulmonary hypertension.(178, 288) This may result in reduced cleavage of ultra-large VWF by ADAMTS13 and coupled with reduced ADAMTS13 levels a prothrombotic state predisposing to, or causing progression of CTEPH. Specific ADAMTS13 activity was increased in CTEPH, which may be a consequence of an increased ratio due to both ADAMTS13 activity *and* antigen levels being reduced, but antigen levels proportionally more so. Alternatively, an increased specific ADAMTS13 activity level may reflect an increased conformational activation of ADAMTS13 by its substrate VWF, due to the altered ADAMTS13:VWF ratio in CTEPH patients.(289)

Further evidence of the role of ADAMTS13 in the pathobiology of thrombosis comes from animal models. ADAMTS13 deficiency increases the infarct size in ischaemic stroke and myocardial infarction murine models and this can be attenuated by recombinant human ADAMTS13 (rhADAMTS13).(290, 291) Furthermore, rhADAMTS13 decreases fibrotic remodelling in a left ventricular pressure overload murine model, which may better reflect

elements of the chronic pathological features in CTEPH.(292) Whilst right ventricular and vascular remodelling occur in CTEPH, right ventricular hypertrophy is not a feature of CTED, in which ADAMTS13-VWF dysregulation also occurs.(276) In ischaemic stroke and myocardial infarction murine models, the cerebral and myocardial injuries induced by ischaemia are VWF-dependent and partially mediated by inflammation.(61, 293)

Inflammation has been linked to CTEPH pathogenesis and some chronic infections are associated risk factors.(294) PEA specimens contain inflammatory cells that correlate with circulating inflammatory markers, which are also increased in CTEPH.(46) The ADAMTS13-VWF axis is abnormal in acute and chronic inflammatory conditions and is posited as a unifying link between inflammation and thrombosis.(286, 295) ADAMTS13 deficiency results in increased leucocyte rolling and adhesion and an increased neutrophil recruitment to the infarcted area in stroke models.(290, 295) Recombinant ADAMTS13 reduces inflammation and platelet recruitment in a left ventricular overload model.(292) In CTEPH, reduced ADAMTS13 would be expected to result in increased inflammation and increased platelet recruitment. No correlation was observed between systemic markers of inflammation and ADAMTS13 or VWF levels suggesting the ADAMTS13-VWF imbalance is not secondary to systemic inflammation. There may still be an interaction with local inflammation in the pulmonary arteries in CTEPH.

4.3.4 ADAMTS13-VWF and ABO

ABO blood groups are associated with CTEPH with an over-representation of the non-O blood group.(68) Genetic variation in *ABO* has also been associated with ischaemic stroke, coronary artery disease and venous thromboembolism.(70, 258) The proposed mechanism of this association has been via VWF plasma levels, which are 25% higher in non-O individuals.(121) VWF is increased in some non-O groups within CTEPH however, VWF is still significantly higher in the CTEPH O group compared with healthy controls. This implies there are other causes of increased VWF and conversely, *ABO* may have additional effects in CTEPH. This would be consistent with studies in VTE where *ABO* remains an independent risk factor after adjusting for VWF levels.(115) *ABO* is a pleiotropic locus and may have alternative functional effects in CTEPH including

mediating pathways involved in inflammation and angiogenesis.(232) Inadequate angiogenesis with a paucity of neovessels and failure to recanalise obstructed vessels has been implicated in CTEPH pathobiology.(45, 46) Interestingly, ADAMTS13 can promote angiogenesis in endothelial cells and therefore, a reduced ADAMTS13 in CTEPH may result in inadequate angiogenesis.(296) Furthermore, in stroke models, ADAMTS13 controls key steps of vascular remodelling and rhADAMTS13 enhances ischaemic neovascularisation.(297)

4.3.5 ADAMTS13 protein quantitative trait loci in CTEPH

A protein quantitative trait loci (rs3739893) was identified in the *C9orf96* gene (~8kb 5' of the *ADAMTS13* gene) that is associated with ADAMTS13 protein levels and has been described in two previous studies.(201, 231) In a GWAS of ADAMTS13 antigen levels in a healthy cohort, this SNP is significantly associated with a similar effect size ($\beta = -22.3\%$). The rs3739893 SNP is highly correlated ($R^2=0.867$, $p<0.001$) with a missense variant (rs41314453) that is the most significant association in an ADAMTS13 activity GWAS.(231) Whilst this suggests that ADAMTS13 protein is genetically regulated, this SNP only accounts for a small amount of variance (~8%) in ADAMTS13 protein levels and is not associated with CTEPH disease risk. ADAMTS13 rare genetic variants have been associated with VTE, which would not have been detected with our minor allele frequency threshold of 1%.(298) Future studies could examine if rare *ADAMTS13* genetic variants are associated with ADAMTS13 protein in CTEPH and if the frequency is different from VTE.

4.3.6 Strengths and limitations

A strength of this study is that it investigated the ADAMTS13-VWF axis in a spectrum of thromboembolic disease from acute PE to chronic thromboembolic disease with and without pulmonary hypertension. The study contains a large sample of well characterised CTEPH patients, who have been phenotyped in expert tertiary centres. ADAMTS13-VWF imbalance does not occur in PE when assessed by multivariable regression. This is consistent with the largest study of ADAMTS13 in VTE, which showed it was not reduced overall.(195) Whilst the study may have been underpowered to detect smaller magnitude

changes in the PE group, ADAMTS13-VWF dysregulation was observed in CTED, a group with a similar sample size. This raises an intriguing possibility, that there are differences in the ADAMTS13-VWF axis in the spectrum of thromboembolic disease ([Section 7.3](#)).

Immunohistochemistry demonstrated that ADAMTS13 is expressed in vascular endothelial cells in control, CTEPH and PEA neovessel samples. However, a limitation was the inability to comprehensively quantify ADAMTS13 in these tissues and establish if the expression differed. Semi-quantitative methods supported by computer-assisted image analysis could be used to assess the expression of the ADAMTS13 antigen between different samples.(229) Alternative methods could be used including quantitative reverse transcription polymerase chain reaction (qRT-PCR), a means of quantifying gene expression by converting ribonucleic acid (RNA) to complementary DNA (cDNA) using reverse transcriptase followed by amplification of specific DNA targets by PCR.(299) Transcriptomics (RNA-sequencing) could be used to quantify ADAMTS13 in different tissues, again by converting messenger RNA (mRNA) to cDNA followed by high-throughput sequencing and alignment with a reference genome or reference transcripts.(300)

The areas for future research for investigating the ADAMTS13-VWF axis in CTEPH are discussed in [Section 7.2](#) and the potential implications for clinical practice are explored in [Section 7.3](#).

In summary, the ADAMTS13-VWF axis is dysregulated in CTEPH and this is unrelated to pulmonary hypertension, disease severity or systemic inflammation. This implicates the ADAMTS13-VWF axis in CTEPH pathogenesis.

5 CTEPH phenotype - genotype associations

5.1 Introduction

This chapter explores additional phenotype-genotype associations by utilising deeply phenotyped data ([Section 5.2.1](#)).

The *ABO* gene locus was the most significant association in the CTEPH GWAS ([Chapter 3](#)). The *F11* locus was a putative association that was significant in the discovery cohort. Genetic risk scores have been developed for venous thromboembolism, whereby disease risk is increased by each additional risk allele.^(301, 302) Furthermore, the combination of factor V Leiden and non-O blood group has a supra-additive effect on VTE risk.⁽¹¹⁵⁾ The effect of a combination of *ABO* and *F11* risk alleles on CTEPH risk is explored in [Section 5.2.2.1](#).

The absence of genetic associations in the CTEPH GWAS may also be revealing important insights into the pathobiology of CTEPH. The majority of CTEPH patients have a preceding VTE (three quarters have a PE and half have a DVT), a polygenic disease with known SNP associations ([Section 1.5.5](#)). In [Section 5.2.2.2](#), the loci associated with VTE are examined in the CTEPH case-control GWAS to identify differential associations. Abnormalities in haemostasis and fibrinolysis are implicated in the pathobiology of CTEPH and patients with CTEPH are treated with anticoagulation, predominately the drug warfarin. Previous GWASs have identified genetic loci that are associated with warfarin metabolism.⁽²³⁶⁾ In [Section 5.2.2.3](#), these loci are examined to establish whether inadequate anticoagulation due to genetic variants related to warfarin metabolism are associated with CTEPH.

ABO is the most significant genetic association in CTEPH however, the pathophysiological consequences of this association are unclear. Evaluation of the ADAMTS13-VWF axis in [Chapter 4](#) suggested that *ABO* may be exerting additional effects ([Section 4.3.4](#)). Whilst *ABO* is associated with CTEPH in a case-control GWAS, its role in CTEPH disease severity and outcomes has not been

investigated. Potential associations between *ABO* and CTEPH haemodynamics and survival are examined in [Section 5.2.3.1](#) and [Section 5.2.3.2](#) respectively.

CTEPH is a heterogenous disease with a spectrum of pulmonary arterial disease distributions and disease severity. CTEPH can occur in different anatomical distributions from the central, proximal pulmonary arteries to the distal vasculature. Distal and proximal CTEPH have been associated with different risk factors that may reflect differing pathobiological mechanisms.(11, 78) Furthermore, patients can have a substantial amount of chronic thromboembolic disease but differing degrees of pulmonary hypertension and right ventricular adaptation.(237) Pathobiological mechanisms or adaptive processes related to disease distribution and severity may have genetic associations that would not necessarily be captured by a case-control analysis. Separate GWAS analyses investigating the common variant associations for CTEPH disease distribution and haemodynamics are investigated in [Section 5.2.3.3](#) and [Section 5.2.3.4](#) respectively.

The aims of this chapter were to investigate:

1. The effect of combining the *ABO* and *F11* risk alleles on CTEPH risk
2. The differential genetic associations between CTEPH, VTE and warfarin metabolism
3. The effect of genetic *ABO* groups on CTEPH disease severity and survival
4. The genetic associations of CTEPH disease distribution and haemodynamics by performing separate GWAS analyses

5.2 Results

5.2.1 Data capture, QC and missingness

To obtain deeply phenotyped CTEPH data for current and future analyses, systematic data extraction and QC were performed as described in [Section 2.3.1.1](#). The additional phenotype - genotype analyses in this chapter predominately contain phenotype data compiled at a single-centre (Papworth) with the highest recruitment. The rationale for this approach is described in [Section 2.3.1](#). The minimal dataset (age, sex and CTEPH disease distribution) missingness varied widely between centres ([Table 5.1](#)).

There were over 200 parameters available for 619 patients from Royal Papworth Hospital included in the GWAS analysis, following QC exclusions. This included the 182 parameters detailed in [Table 2.2](#) (Material and Methods) and additional variables from echocardiograms, radiological investigations, co-morbidities and demographic details.

Extensive QC was performed on the separate datasets as described in [Section 2.3.1.1](#). As an exemplar, the QC steps for the haemodynamic parameters (n=30 variables) from right heart catheterisations at Royal Papworth Hospital (n=3366 individuals in the starting dataset) will be described.

	Bad Nauheim	Imperial	Leuven	Papworth	San Diego	Vienna	Total
Age	0 (0)	9 (15)	0 (0)	18 (3)	0 (0)	143 (100)	170 (14)
Sex	0 (0)	7 (12)	0 (0)	18 (3)	0 (0)	0 (0)	25 (2)
Disease distribution	2 (1)	59 (100)	13 (9)	22 (4)	50 (50)	66 (46)	212 (17)

Table 5.1 Missingness of variables from different recruitment centres in the GWAS minimal dataset following QC removals

n=1250 included in the GWAS analysis (see [Figure 3.7](#) for the numbers included from each centre).

Data were harmonised by ensuring the correct data type (e.g. continuous data) and units were present. Physiologically impossible values were removed (e.g. negative values) and then physiological rules were applied (e.g. cardiac index is always less than cardiac output) with further outlier removal. Test dates were standardised and merged with other datasets to obtain the baseline (closest to the time of diagnosis) values. Additional rules for test dates (e.g. baseline haemodynamic parameters had to be prior to surgery for those undergoing PEA) were then applied to select appropriate data for the downstream analyses. The importance of systematic and reproducible QC steps is highlighted by the difference in the density plots pre- and post-QC shown in [Figure 5.1](#), which are essential for determining the most suitable statistical analysis.

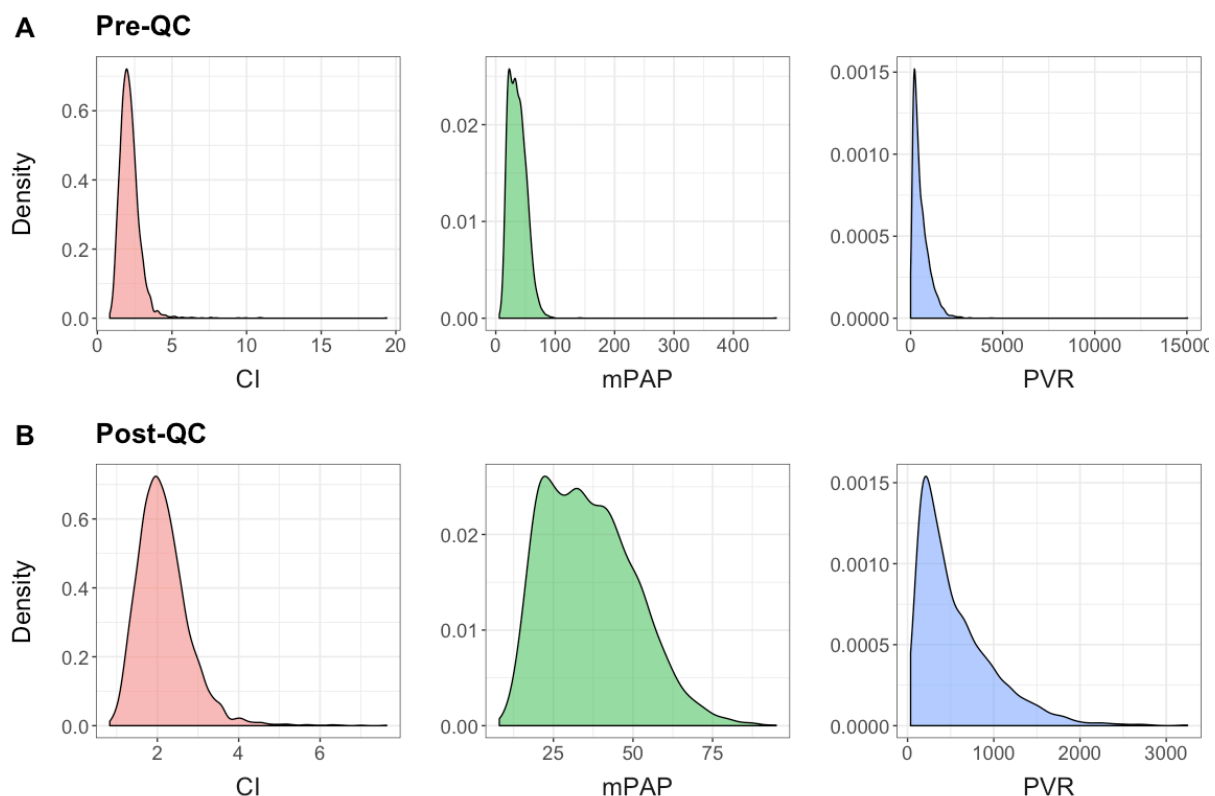


Figure 5.1 Density plots of selected haemodynamics pre- and post-QC

A Density plots of CI (cardiac index), mPAP (mean pulmonary artery pressure) and PVR (pulmonary vascular resistance) prior to QC and **B** following QC steps (n=50-100 removals). Starting dataset n=3366 (all right heart catheterisations from Royal Papworth Hospital). Subsequently, this dataset was used for the GWAS analysis ([Section 5.2.3](#)).

5.2.2 Additional GWAS case-control analysis

5.2.2.1 *ABO* and *F11* risk alleles and CTEPH

The risk of CTEPH is increased as the number of risk (effect) alleles at the significant loci in *ABO* (chr9) and the putative association in *F11* (chr4) increase (**Figure 5.2**). CTEPH risk is greater for patients with one risk allele at *ABO* (rs2519093-T allele; OR (95% CI): 2.83 (1.91-4.22)) compared with one risk allele at *F11* (rs2036914-C allele; OR (95% CI): 1.48 (1.09-2.03)). The combination of one risk allele at both the *ABO* and *F11* loci has an additive effect (3.46 (2.49-4.85)), which continues with each additional risk allele (**Figure 5.2**). Patients with 2 risk alleles at both the *ABO* and *F11* loci (4 risk alleles in total) have the highest risk of CTEPH (7.43 (3.72-15.5)), albeit with a wide 95% confidence interval due to the small group size. Only 7% (n=87) CTEPH patients have no risk alleles compared with 16% (n=243) of healthy controls.

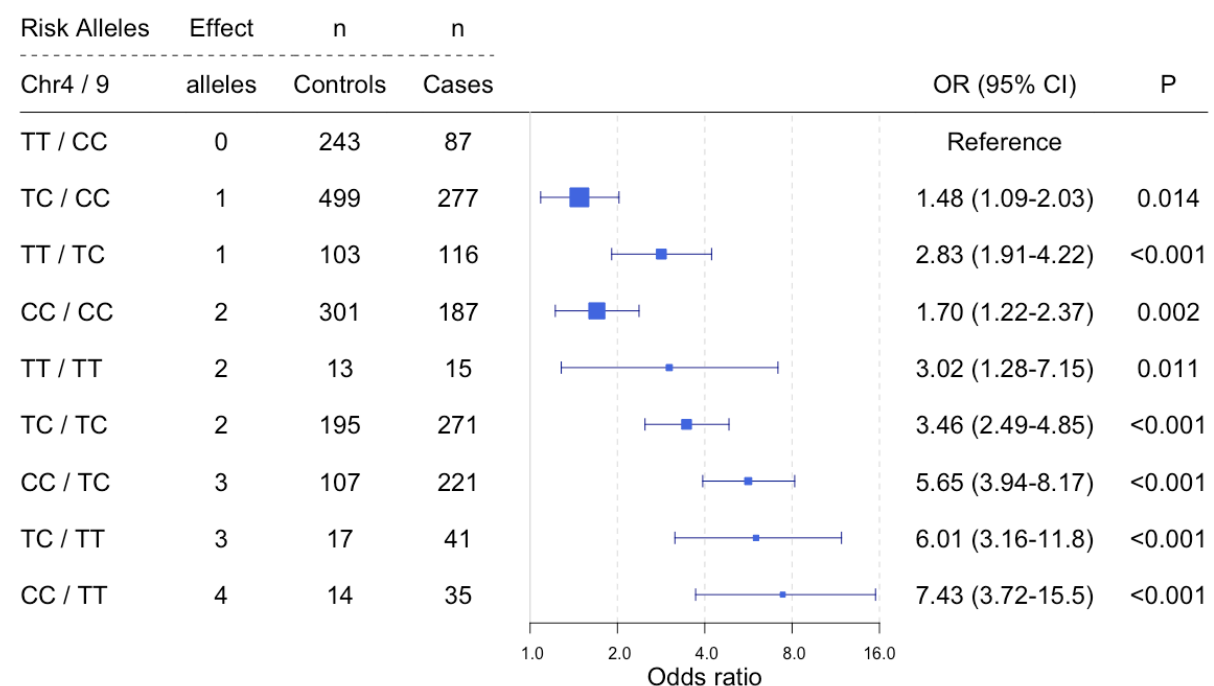


Figure 5.2 *ABO* and *F11* risk alleles and CTEPH

Odds ratios for CTEPH (with respect to healthy controls) in different *ABO* and *F11* risk allele groups calculated using logistic regression adjusted for 5 ancestry informative principal components. The risk allele for rs2519093 (Chr9, *ABO*) is T and the risk allele for rs2036914 (Chr4, *F11*) is C. Each additional risk allele generally

confers an additive effect on disease risk. There are some combinations such as increasing from 1 to 2 *F11* risk alleles (TC / CC *versus* CC / CC) that may not confer an additive effect (OR 1.48 *versus* 1.70). However, their 95% CI (1.09-2.03 *versus* 1.28-2.37) could still be consistent with an additive effect and moreover the trend of increasing risk alleles from 0 to 4 is additive.

5.2.2.2 Venous thromboembolism genes in CTEPH

Of the 9 loci (in 8 gene regions) that are significantly associated with VTE in a GWAS meta-analysis (7,507 VTE cases and 52,632 healthy controls), only the *ABO* locus is associated with CTEPH (**Figure 5.3** and **Table 5.2**).⁽⁷⁰⁾ The most significant *ABO* locus SNP in the VTE GWAS (rs529565) is also highly associated with CTEPH (OR (95% CI): 1.9 (1.77-2.02), $p=4.42 \times 10^{-23}$) and moderately correlated with the lead *ABO* SNP association in the CTEPH GWAS ($R^2=0.433$, European (non-Finnish) 1000Genomes phase 3 data). The putative SNP association in the *F11* locus from the CTEPH GWAS (rs2036914) is only in moderate-low correlation with the VTE *F11* SNP association (rs4253417, $R^2=0.357$).

There are no significant CTEPH associations for the other VTE associated SNPs (*F5*, *FGG*, *TSPAN15*, *SLC44A2* and *PROCR*) (**Figure 5.3** and **Table 5.2**). There are several non-significant putative association signals particularly in the *F5* (lead SNP rs2009814, effect allele T, OR (95% CI) = 1.36 (1.22-1.50), $p=2.60 \times 10^{-5}$), *FGA-FGB-FGG* (lead SNP rs13130318, effect allele T, OR (95% CI) = 1.41 (1.28-1.54), $p=4.45 \times 10^{-7}$) and *F11* (described in **Section 3.2.3**) loci (**Figure 5.3**).

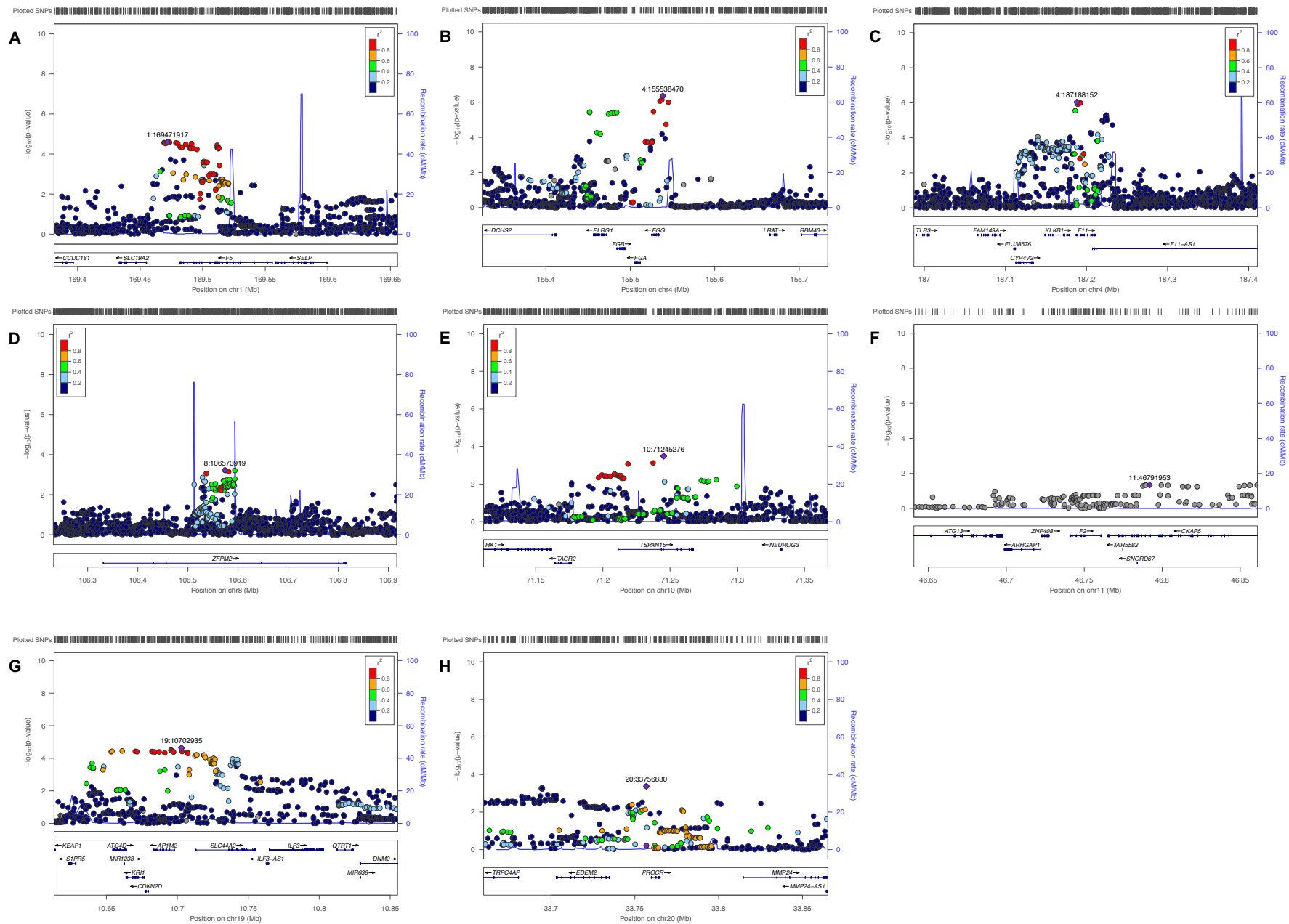


Figure 5.3 VTE associated loci in the CTEPH GWAS

Regional association plots (LocusZoom) from the CTEPH GWAS (1250 CTEPH patients, 1492 healthy controls and 7,675,738 SNPs) focusing on the significant loci that have previously been described for VTE.⁽⁷⁰⁾ The significant association in *ABO* (chr9) detailed in [Chapter 3](#) is not shown. The lead SNPs displayed in the LocusZooms are not necessarily the same as in [Table 5.2](#).

			VTE			CTEPH			
rsID	CHR:POS_EA/NAE	GENE	EAf	OR (95% CI)	<i>p</i>	EAf	OR (95% CI)	<i>p</i>	INFO
rs4524	1:169511755_T/C	F5	0.736	1.20 (1.14–1.26)	2.65e-11	0.775	1.33 (1.19-1.46)	4.77e-05	1.000
rs6025	1:169519049_T/C	F5	0.033	3.25 (2.91–3.64)	1.1e-96	0.047	1.02 (0.677-1.36)	9.08e-01	0.588
rs2066865	4:155525276_A/G	FGG	0.244	1.24 (1.18–1.31)	1.03e-16	0.605	1.38 (1.21-1.55)	1.92e-04	0.525
rs4253417	4:187199005_C/T	F11	0.405	1.27 (1.22–1.34)	1.21e-23	0.094	0.791 (0.571-1.01)	3.65e-02	0.681
rs529565	9:136149500_C/T	ABO	0.354	2.29 (1.75–2.99)	1.73e-09	0.463	1.9 (1.77-2.02)	4.42e-23	0.976
rs78707713	10:71245276_T/C	TSPAN15	0.878	1.15 (1.10–1.21)	1.65e-08	0.907	1.45 (1.25-1.66)	3.29e-04	0.852
rs1799963	11:46761055_A/G	F2	0.010	1.20 (1.13–1.27)	3.48e-09	NA	NA	NA	0.030
rs2288904	19:10742170_G/A	SLC44A2	0.785	1.28 (1.19–1.39)	5.74e-11	0.817	1.32 (1.18-1.47)	2.24e-04	1.000
rs6087685	20:33777612_C/G	PROCR	0.302	1.19 (1.12–1.26)	1.07e-09	0.192	0.938 (0.769-1.11)	4.55e-01	0.859

Table 5.2 VTE associated loci in the CTEPH GWAS

The VTE associations shown in the table are from a GWAS meta-analysis and the associations for CTEPH are from the CTEPH GWAS ([Chapter 3](#)).⁽⁷⁰⁾ The rs1799963 SNP (*F2* gene locus) was not analysed in the CTEPH GWAS as it was filtered out due to a low effect allele frequency (< 1%) and low information score (INFO < 0.5). The allele frequencies for the *FGG* and the *F11* gene locus SNPs differ markedly between CTEPH, VTE and a European (non-Finnish) reference population (0.239 and 0.403 respectively; not shown in table). This is likely to be related to poorly imputed SNPs reflected by a lower INFO score. rsID (reference SNP cluster ID), CHR (chromosome), POS (Base position, GRCh37 genome build), EF (effect allele), NEA (non-effect allele), GENE (nearest gene of the SNP, from ANNOVAR), EAF (effect allele frequency of VTE/CTEPH patients), INFO (information score, imputation quality), OR (odds ratio), *p* (*p*-value).

The *F5* SNP (rs6025; the factor V Leiden variant) is the most significant association with the highest odds ratio in the VTE GWAS (OR (95% CI): 3.25 (2.91-3.95), $p=1.1 \times 10^{-96}$) but is not significantly associated with CTEPH (OR (95% CI): 1.02 (0.67-1.36), $p=0.908$) The lack of VTE associations in the CTEPH GWAS may be due to differing genetic aetiology or a lack of power to detect associations. To address this, power calculations were performed for the *F5* locus ([Figure 5.4](#)). With the current CTEPH GWAS sample size (cases=1250, controls=1492) there is 100% power to detect an rs6025 association assuming a disease prevalence of 0.00003, a disease allele frequency of 0.033, an OR of 3.25 (from [Table 5.2](#)) and an additive genetic model. However, for the other significant SNP in *F5* (rs4524) over 15,000 cases and controls would be required to achieve a power of 80% (assuming an OR of 1.2 and disease allele frequency of 0.736). This suggests that we are adequately powered to detect a lack of association in rs6025, but under-powered to detect some other VTE associations.

5.2.2.3 SNPs associated with warfarin metabolism in the CTEPH GWAS

The 4 loci (in 3 gene regions: *CYP2C9*, *VKORC1*, *CYP4F2*) associated with warfarin metabolism are not associated with CTEPH in the GWAS ([Figure 5.5](#) and [Table 5.3](#)).⁽²³⁶⁾ Most notably, the SNP in *VKORC1* (rs9923231) that explains ~30% of the warfarin dose variance does not increase the risk of CTEPH (OR (95% CI): 0.872 (0.751-0.993), $p=0.027$).

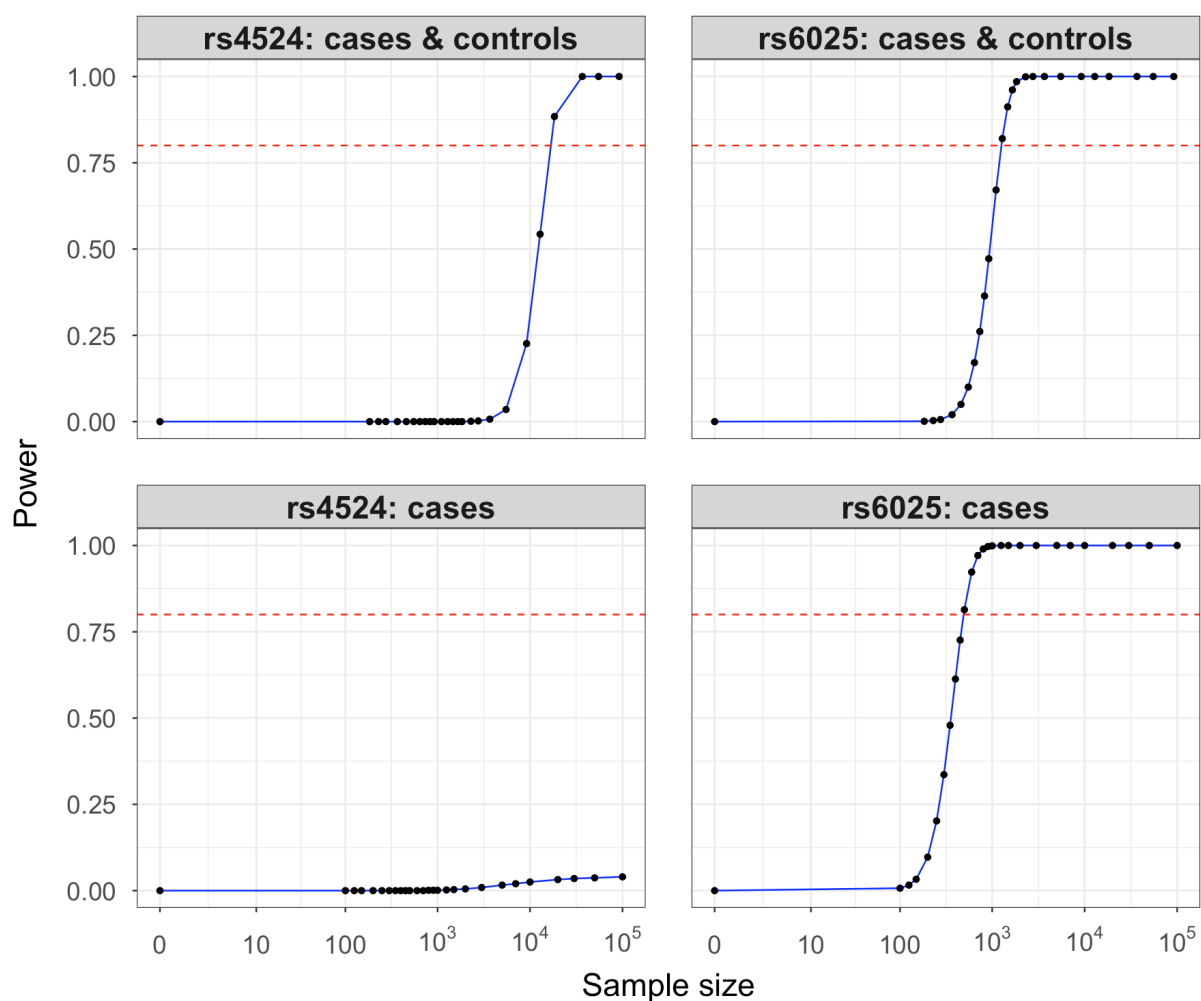


Figure 5.4 Power calculations for the *F5* VTE associated loci

Power calculations for the SNPs rs4524 and rs6025, which are the significant *F5* gene locus associations in the VTE GWAS. The power calculations assume an allele frequency of 0.736 and OR of 1.2 for rs4524 and an allele frequency of 0.033 and OR of 3.25 for rs6025 (taken from the VTE GWAS results shown in [Table 5.2](#)). A disease prevalence of 0.00003 and an additive genetic model were used for the calculations. In the top panels power calculations were performed using disease cases (n=1250) and healthy controls (n=1492), whereas in the bottom panels only disease cases (n=1250) were included in calculations.(303) Power calculations were performed using the GAS online power calculator.(203)

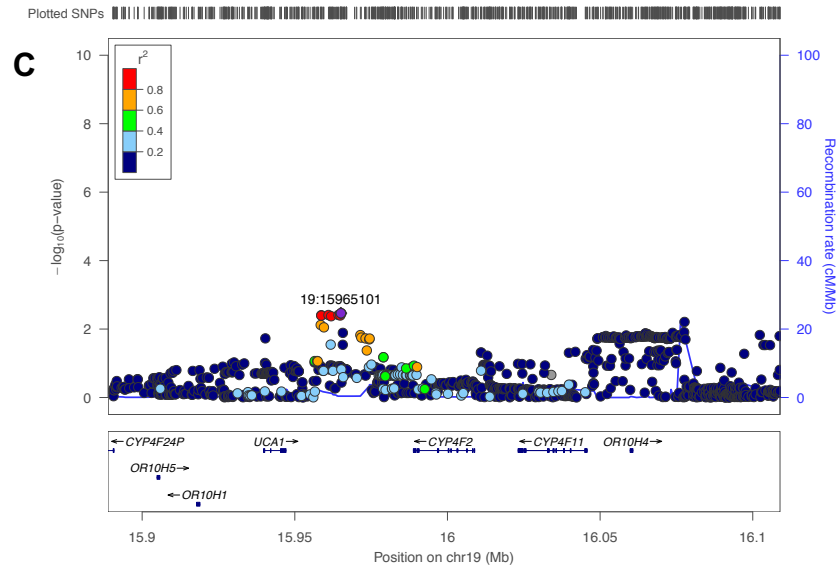
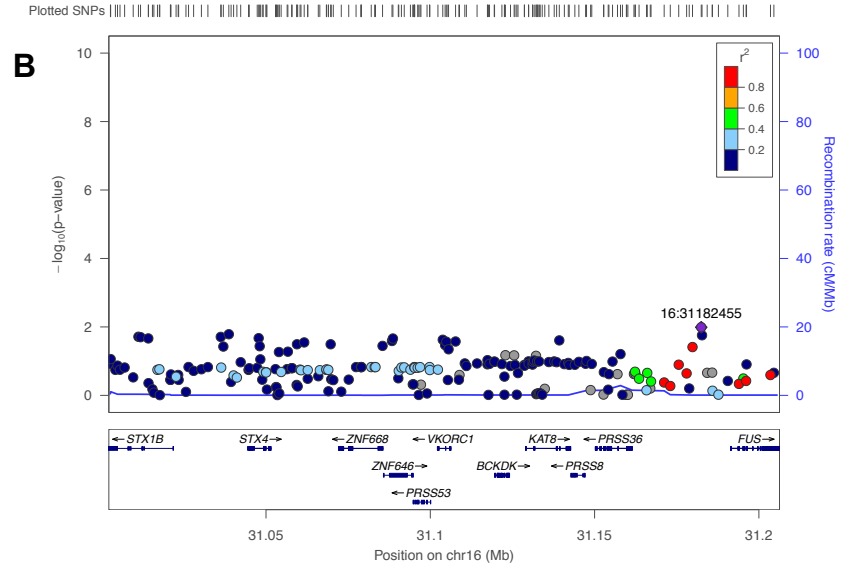
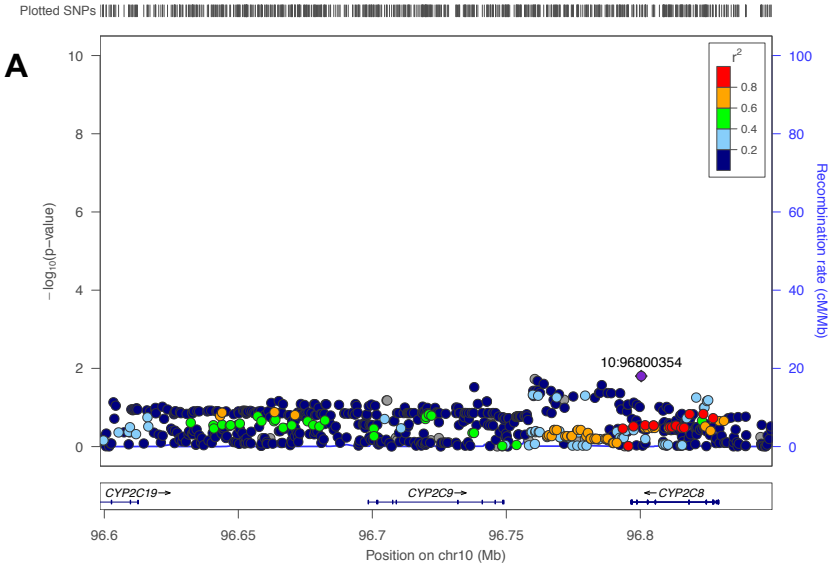


Figure 5.5 Warfarin metabolism associated loci in the CTEPH GWAS

Regional association plots (LocusZoom) from the CTEPH GWAS (1250 CTEPH patients, 1492 healthy controls and 7,675,738 SNPs) focusing on the significant loci that have previously been described for warfarin metabolism.(236) The variants annotated in the plots are the most significant ones for that region and do not correspond to the SNPs associated with warfarin metabolism described in more detail in [Table 5.3](#).

rsID	CHR:BP_EA/NEA	GENE	EAf_A	EAf_U	EAf_REF	OR (95% CI)	p	INFO
rs1799853	10:96702047_C/T	CYP2C9	0.250	0.243	0.1266	0.958 (0.807-1.11)	0.581	0.803
rs1057910	10:96741053_C/A	CYP2C9	0.0604	0.0694	0.0626	0.835 (0.596-1.07)	0.129	1.000
rs9923231	16:31107689_T/C	VKORC1	0.360	0.383	0.368	0.872 (0.751-0.993)	0.027	1.000
rs2108622	19:15990431_T/C	CYP4F2	0.299	0.302	0.287	1.04 (0.91-1.16)	0.583	1.000

Table 5.3 Warfarin metabolism associated loci in the CTEPH GWAS

The SNPs associated with warfarin metabolism were defined in a GWAS of ~1500 patients taking warfarin from a Swedish population.(236) It was not clear from the *Takuchi et al*, study what the effect allele (EA) for each SNP was, so this has been defined as the minor allele. The allele frequencies for rs1799853 (*CYP2C9*) differ between the CTEPH GWAS and a reference population, which may be due to the lower quality imputation of this SNP. EAF_A (effect (minor) allele frequency of affected CTEPH patients),

EAF_U (effect (minor) allele frequency of unaffected healthy controls), EAF_REF (effect (minor) allele frequency of reference, 1000 genomes phase 3 European (non-Finnish) populations). The additional column headings are described in [Table 5.2](#).

5.2.3 Additional phenotype-genotype associations

5.2.3.1 *ABO* and CTEPH disease severity

Haemodynamic parameters from right heart catheterisation performed at the time of CTEPH diagnosis are a marker of disease severity. Higher mean pulmonary arterial pressure (mPAP) / pulmonary vascular resistance (PVR) and lower cardiac index (CI) reflect more severe disease. The effect of the alleles for the significant chr9 locus (*ABO*) and the putative chr4 locus (*F11*) in the CTEPH GWAS on baseline haemodynamics was investigated. There was no difference in the baseline haemodynamics of all CTEPH patients with risk alleles in *ABO* or *F11* when analysed separately or in combination ([Figure 5.6](#)).

Haemodynamics do not vary with inferred genetic *ABO* groups (described in [Section 2.1.8](#)) for CTEPH patients ([Figure 5.7A-C](#)). Additional markers of CTEPH disease severity (six-minute walk distance and WHO functional class) also do not vary with genetic *ABO* groups ([Figure 5.7D](#) and [5.7E](#)).

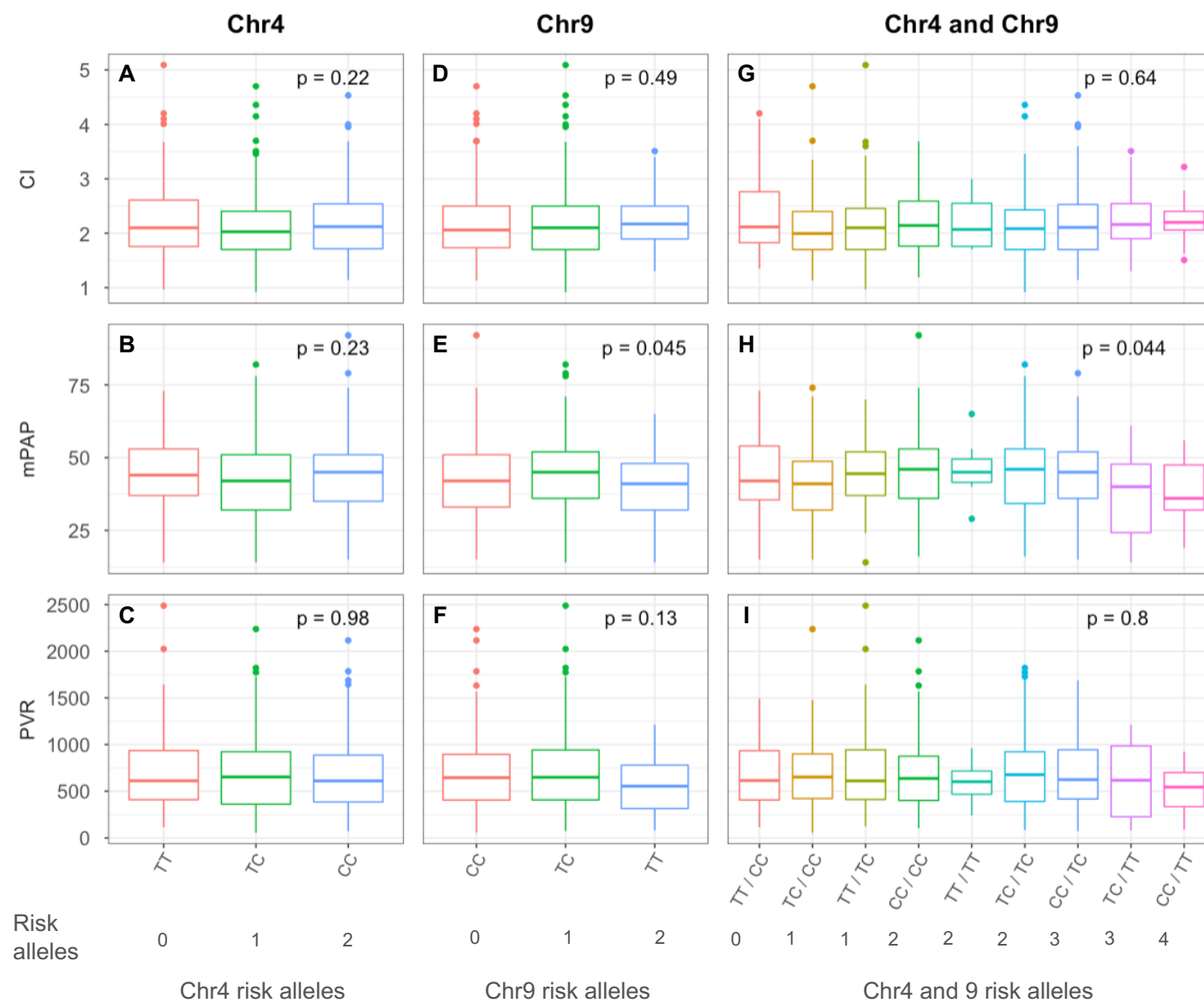


Figure 5.6 The effect of risk (effect) alleles for Chr4 (*F11*) and Chr9 (*ABO*) on haemodynamics

CTEPH patients were subdivided by their number of risk (effect) alleles from the *ABO* (risk allele T) and putative *F11* (risk allele C) associations in the CTEPH GWAS. Group differences in haemodynamics (CI, mPAP and PVR) were then assessed using the Kruskal-Wallis test. The number of CTEPH patients with available data for CI, mPAP and PVR were: 551, 625 and 610 respectively. The nominally significant *p*-values in **E** and **H** are no longer significant when adjusted for multiple testing. CI (cardiac index, L/min/m²), mPAP (mean pulmonary arterial pressure, mmHg), PVR (pulmonary vascular resistance, dynes/sec/cm⁻⁵).

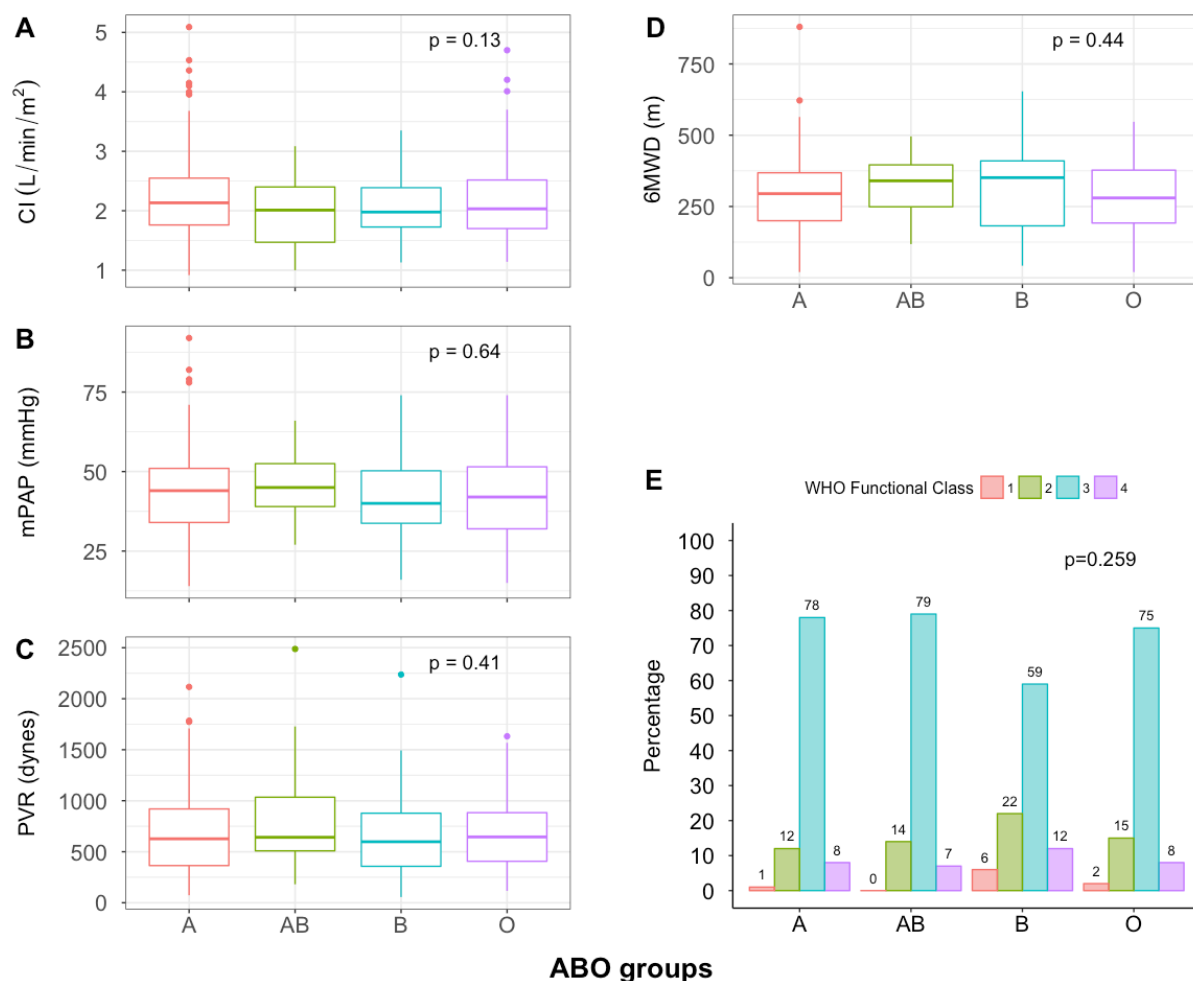


Figure 5.7 Genetic ABO groups and CTEPH disease severity

CTEPH patients were subdivided into their inferred genetic ABO group (A, AB, B and O). Group differences in haemodynamics (CI, mPAP and PVR) and 6mwd were assessed using the Kruskal-Wallis test and the Cochran-Armitage test for WHO functional class. The number of CTEPH patients with available data for CI, mPAP, PVR, 6mwd and WHO functional class were: 535, 604, 589, 455 and 525 respectively. The available haemodynamic data differs from [Figure 5.6](#) as a genetic ABO group could not be inferred in all CTEPH patients (see [Section 2.1.8](#)). In [Figure 5.7E](#) the percentage for each group is shown above the bars. CI (cardiac index, L/min/m²), mPAP (mean pulmonary artery pressure, mmHg), PVR (pulmonary vascular resistance, dynes/sec/cm⁻⁵), 6mwd (six-minute walking distance, metres), WHO (World Health Organisation).

5.2.3.2 ABO groups and CTEPH survival

Survival was investigated in CTEPH patients following PEA as this represented the largest group that received the same intervention. Mortality data was only available from Royal Papworth Hospital. Of the 619 CTEPH patients from Papworth included in the CTEPH GWAS analysis, 454 underwent PEA and 421 of these had both mortality data and genetically inferred ABO groups available. There was no difference in survival following surgery between genetic ABO groups in CTEPH (log-rank: $p=0.29$) (Figure 5.8 and Table 5.4). Survival was worse with increasing age (hazard ratio (95% CI) per 1 year increase: 1.04 (1.02-1.06), $p<0.001$) (Table 5.4).

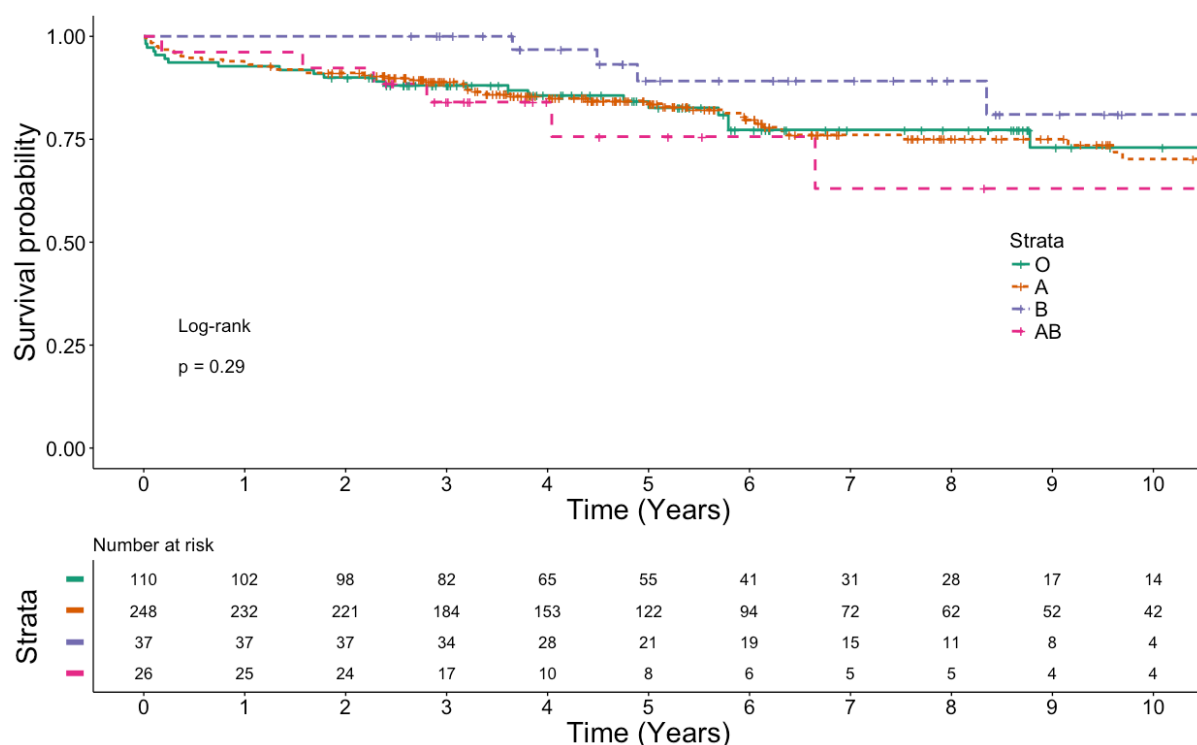


Figure 5.8 CTEPH survival following pulmonary endarterectomy in different ABO groups

Kaplan-Meier survival curves of survival time for post-PEA CTEPH patients stratified by genetic ABO group. Group differences were assessed with a log-rank test.

As multiple variables can influence post-PEA survival in CTEPH, a cox proportional hazards model was constructed.(15) Survival did not vary by genetic ABO groups when adjusting for age, sex and baseline pre-operative disease severity (mPAP) (Table 5.4). The small sample size of some ABO groups (e.g. B and AB) made it difficult to be definitive on their survival effect.

	n	HR	95% CI	p
ABO: O	103	Reference		
ABO: A	237	1.05	0.64 - 1.75	0.836
ABO: B	34	0.70	0.26 - 1.85	0.467
ABO: AB	26	1.89	0.83 - 4.3	0.131
Female	174	Reference		
Male	226	0.94	0.62 - 1.44	0.779
Age	400	1.04	1.02 - 1.06	<0.001
mPAP: Baseline	400	1.01	0.99 - 1.03	0.429

Table 5.4 Cox proportional hazards model of post-PEA survival in CTEPH

Cox proportional hazards model assessing time to death following PEA surgery. There were 400 CTEPH patients (54 removed due to missing variables) and 89 events (deaths) included in the model. Hazard Ratios (HR) are shown with 95% confidence intervals. mPAP was included as a covariate to account for baseline pre-surgical disease severity as it was the haemodynamic measurement with the most data points. Data analysis was performed and models checked using the R packages: `survival` and `survminer`. (249, 250)

5.2.3.3 CTEPH disease distribution GWAS

1035 (83%) of patients included in the CTEPH GWAS had a disease subtype (CTEPH distal / proximal or CTED) recorded. Of the 983 patients with CTEPH, 847 (86%) had a proximal distribution in the pulmonary arteries and 136 (14%) had a distal distribution. The variation in disease sub-type proportions by centre is shown in [Figure 5.9](#).

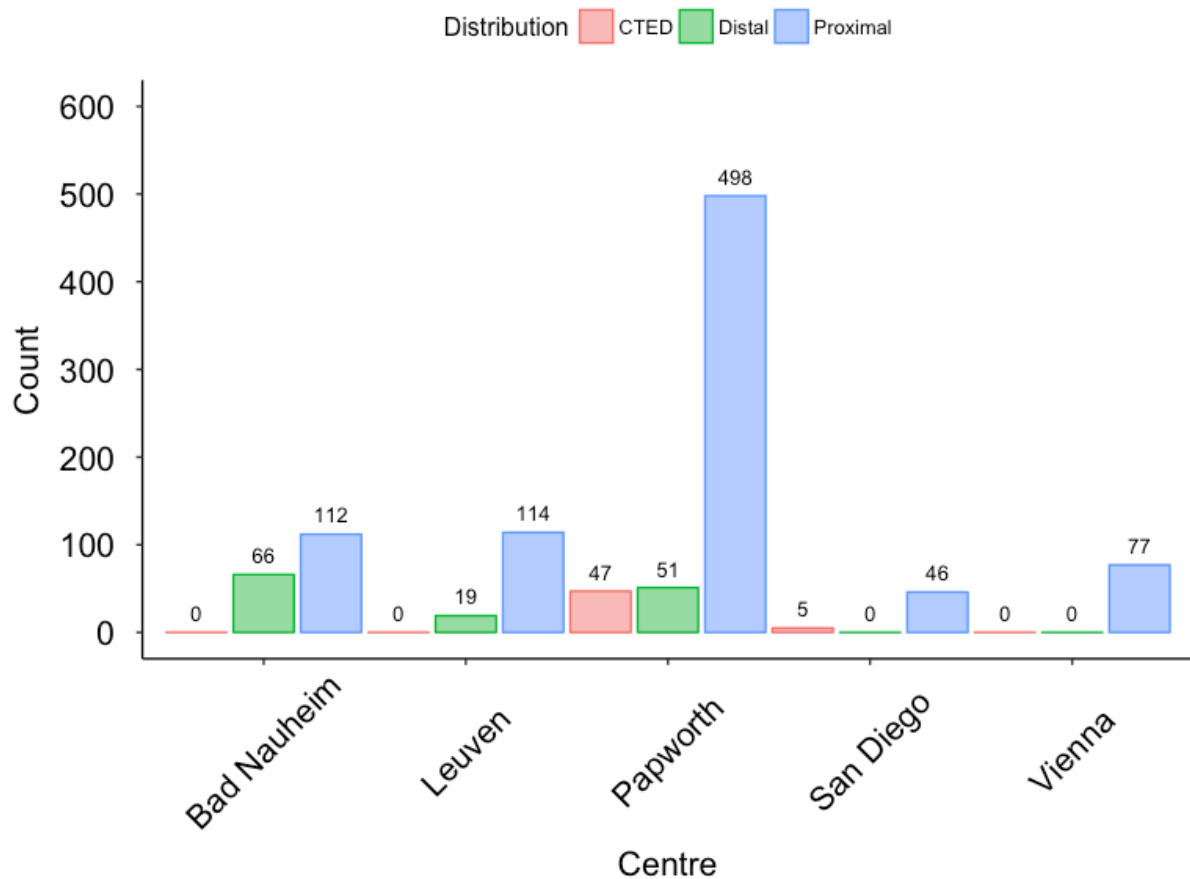


Figure 5.9 CTEPH disease subtypes by recruiting centre

Barplot of the number of CTEPH disease subtype (CTEPH: proximal or distal, and CTED) patients from each centre. 1035 patients had information on disease subtype and 215 were missing these data. The count (n) for each group is shown above the bars. Centres that did not supply data on disease subtype are not included in the plot.

A separate GWAS analysis was performed to investigate if there were genetic associations with CTEPH disease subtype. Imputed genotype dosages were used to test for an association between proximal (n=847) and distal (n=136) CTEPH patients following QC and imputation steps as previously described ([Chapter 3](#)). Association testing was performed as described using logistic regression with post-imputation SNP dosages (n=7,675,738) assuming an additive model and adjusted for 5 principal components, age, sex and centre ([Figure 5.10A, C and D](#)). The additional covariates (age, sex and centre) were included as they can vary between proximal and distal disease.^(11, 78) No locus was genome-wide significant ($p < 5 \times 10^{-8}$). The most

significant SNPs are rs2313920 (OR (95% CI) 0.317 (0.303-0.331), $p=7.63 \times 10^{-8}$) an intronic variant in the *KSR1* (Kinase Suppressor of Ras 1) gene and rs68044424 (OR (95% CI) 4.44 (3.88-5.01), $p=2.15 \times 10^{-7}$) an intergenic SNP in chromosome 11.

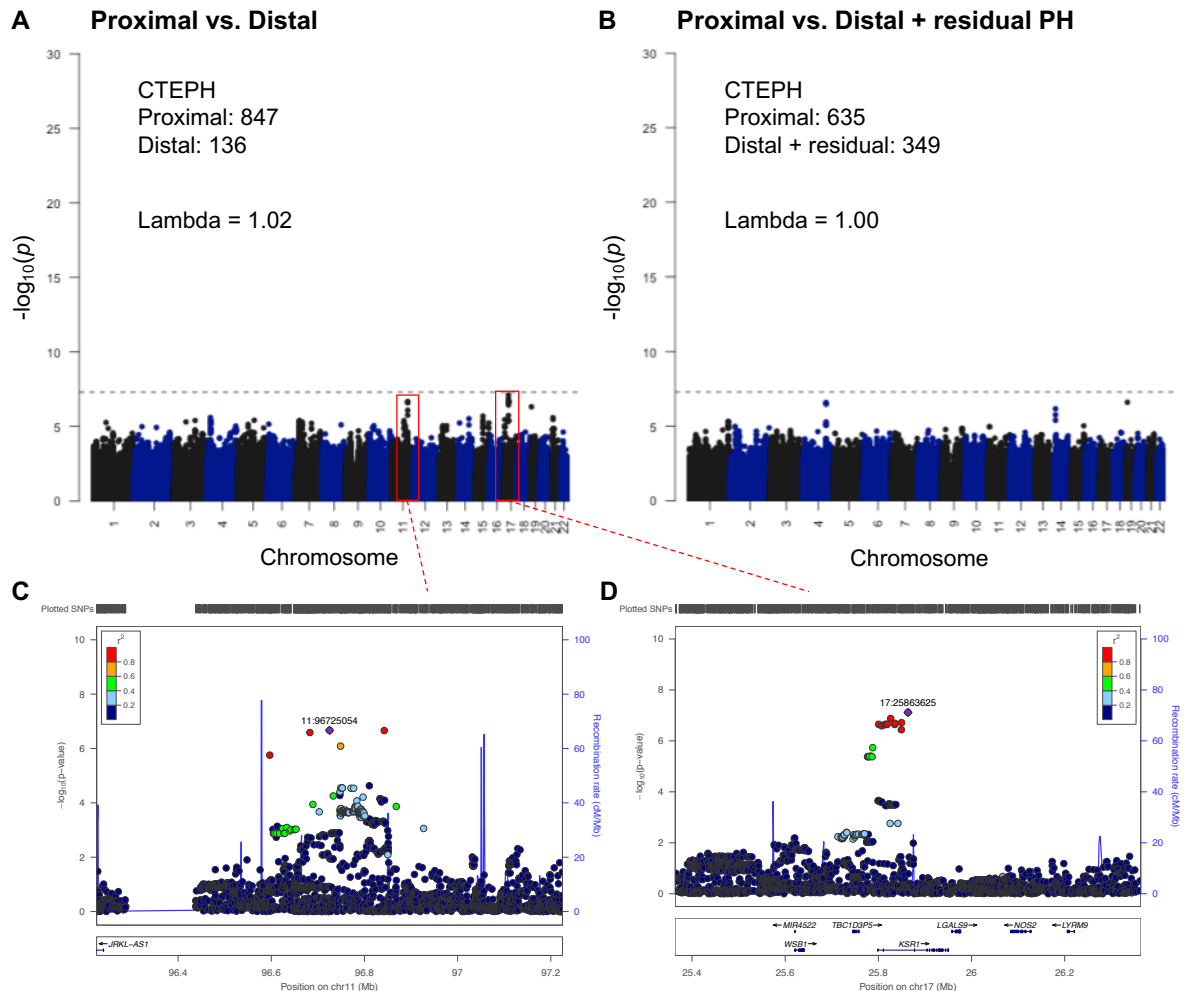


Figure 5.10 Distal/Proximal CTEPH GWAS association testing

Analysis of 136 distal CTEPH, 847 proximal CTEPH and 7,675,738 SNPs. Statistical testing of individual SNPs using allelic dosage (range 0-2) for an association with CTEPH disease subtype was performed using logistic regression assuming an additive genetic model and age, sex and 5 ancestry informative principal components were included as covariates. A p -value of $<5 \times 10^{-8}$ was considered genome-wide significant (grey dotted line). P -values are transformed to a $-\log_{10}$ scale.

A Distal vs. Proximal CTEPH association testing. The most significant associations in chromosome 17 (rs2313920) corresponds to 17:25863625, and in chromosome 11 (rs68044424) to 11:96725054 on the plot.

B Distal (including post-PEA residual pulmonary hypertension) (n=349) vs. Proximal CTEPH (n=635)

C and D Regional association plots of chromosomes 11 and 17 showing the most significant putative associations from [Figure 5.10A](#).

The distal/proximal CTEPH association testing is limited by the small sample size of the distal CTEPH group. Over 50% of patients have persistent pulmonary hypertension following pulmonary endarterectomy.(15, 20) As the majority of proximal thromboembolic material is removed during surgery, the residual pulmonary hypertension may be a consequence of distal vasculopathy. There may be an overlap in the pathobiology of distal CTEPH and post-PEA residual pulmonary hypertension with shared genetic associations. This was explored by performing a GWAS analysis that included post-PEA residual pulmonary hypertension patients (n=213) with distal CTEPH (n=136) in one group (n=349 total) and compared them to proximal CTEPH (n=635) in the second group. Association testing was performed as described for the original distal/proximal CTEPH groups. There were no loci that achieved genome-wide significance ([Figure 5.10B](#)).

5.2.3.4 CTEPH haemodynamics GWAS

Another separate GWAS analysis was performed to investigate if there were genetic associations with CTEPH haemodynamics, which are a marker of disease severity. Haemodynamics (mPAP, CI and PVR) obtained at the time of baseline (diagnostic) right heart catheterisation were utilised in linear regression with post-imputation SNP dosages (n=7,675,738). The model was adjusted for 5 ancestry informative principal components and factors that could affect haemodynamics: age, sex and recruiting centre. No locus was genome-wide significant ([Figure 5.11](#)). The most significant association occurred in the cardiac index GWAS (rs4240181, β (95% CI):-0.148 (-0.150 to -0.146), $p=4.16 \times 10^{-7}$) and was an intronic variant in the *TUSC3* (Tumour Suppressor Candidate 3) gene. Another putative association was the SNP rs145980813 (β (95% CI):-0.340 (-0.191 to -0.489), $p=4.45 \times 10^{-7}$) an intronic variant in *MALRD1* (MAM And LDL Receptor Class A Domain Containing 1).

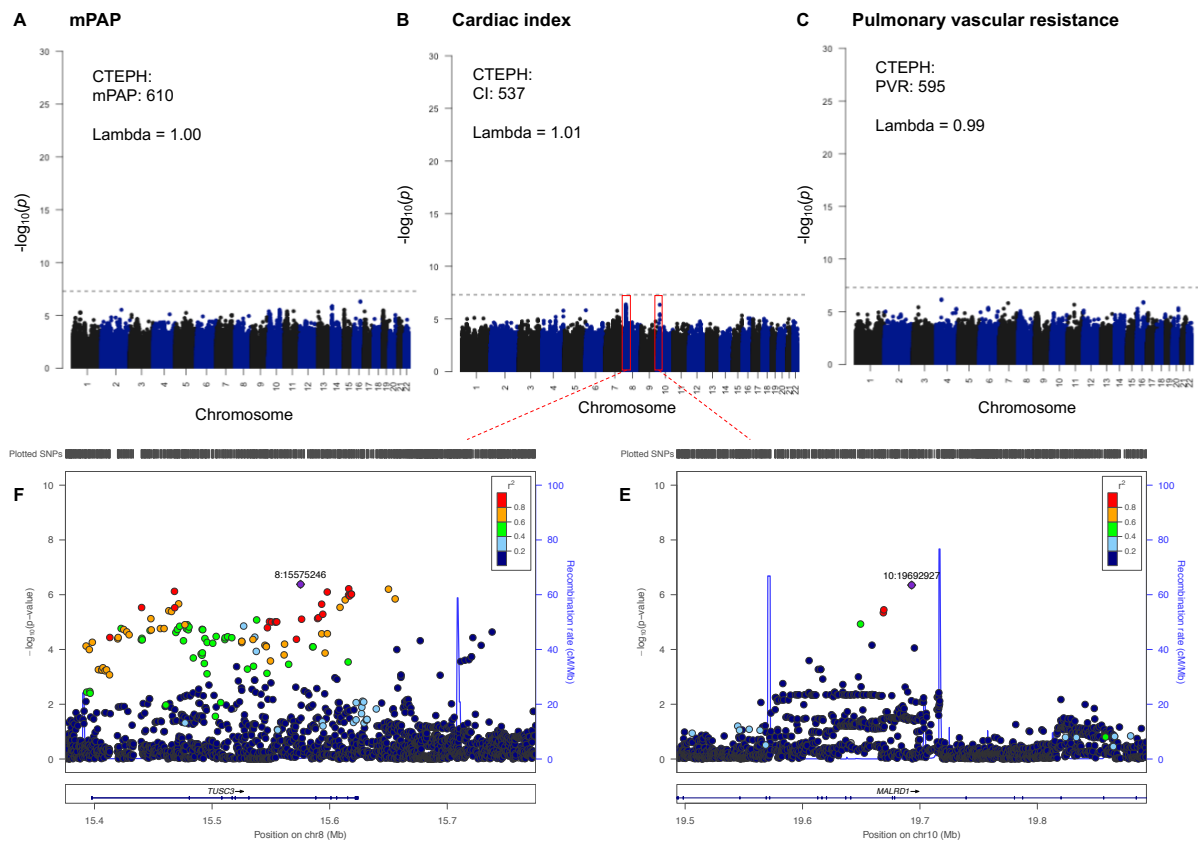


Figure 5.11 CTEPH haemodynamic GWAS association testing

GWAS analysis of baseline (diagnostic) haemodynamics (mPAP, CI and PVR) obtained at right heart catheterisation for CTEPH utilising 7,675,738 SNPs. The number of individuals with mPAP, CI and PVR data available were: 610, 537 and 595 respectively. Statistical testing of individual SNPs using allelic dosage (range 0-2) for association with CTEPH haemodynamics (mPAP, CI or PVR) was performed using linear regression assuming an additive genetic model and age, sex and 5 ancestry informative principal components were included as covariates. As some haemodynamic parameters had skewed distributions, linear models were first constructed without the SNP allelic dosages to evaluate the normality of residuals. Appropriate transformations were then applied if required (square root (PVR), log (CI)) and used in the linear models with the SNP allelic dosages and covariates. A p -value of $<5 \times 10^{-8}$ was considered genome-wide significant (grey dotted line). P -values are transformed to a $-\log_{10}$ scale.

A Mean pulmonary arterial pressure (mPAP)

B Cardiac index (CI)

C Pulmonary vascular resistance (PVR)

D and E Regional association plots for the most significant putative associations in chromosomes 8 and 10 from the cardiac index analysis ([Figure 5.11B](#))

5.3 Discussion

5.3.1 The effect of combining the *ABO* and *F11* risk alleles on CTEPH risk

ABO is the most significant genetic association for CTEPH followed by a putative association in a *F11* locus. Combining risk alleles at these two loci does not result in a supra-additive effect on CTEPH disease risk. Only 7% of CTEPH patients have no risk alleles at *ABO* and *F11* compared with 16% of healthy controls. However, given the common allele frequencies of the *ABO* and *F11* SNPs, these variants are neither necessary nor sufficient to develop CTEPH.

5.3.2 The differential genetic associations between CTEPH, VTE and warfarin metabolism

Given that three quarters of CTEPH patients have had a preceding pulmonary embolism, we may expect similar common variant genetic associations. Of the 9 common genetic loci associated with VTE, only *ABO* is significantly associated in the CTEPH GWAS. This may be due to a relatively small sample size and lack of power to detect associations. However, the CTEPH GWAS is adequately powered to exclude a factor V Leiden variant (rs6025) association. This suggests that either the factor V Leiden variant is not significantly associated with CTEPH, or that an association exists but with a smaller effect size.

The factor V Leiden polymorphism is a missense mutation that results in activated protein C resistance and a procoagulant state (56). It affects ~5% of Caucasians and is much rarer in other ethnic groups.(254) Individuals heterozygous for factor V Leiden have a 3-5 fold increased lifetime risk of VTE, and in homozygotes the risk is increased over 10 fold.(56) Furthermore, 20-25% of patients with a first episode of unprovoked VTE have the factor V Leiden polymorphism.(304) The risk of VTE recurrence in heterozygous carriers of factor V Leiden is only modest (OR ~1.4).(305) Whilst the mutation occurs more frequently in isolated PE than healthy controls (OR ~2), there is an even higher frequency in DVT (OR ~4).(106, 306, 307) The stronger association with DVT than PE is termed the factor V paradox. The reasons for the apparent paradox are unclear, but may relate to a reduced embolization risk conferred by factor V Leiden.(308)

In the CTEPH GWAS, the lack of association with factor V Leiden could be attributable to differences in the CTEPH and VTE GWAS cohorts. All VTE GWAS cohorts have pooled patients diagnosed with DVT, PE or a combination of both.(70) To date, the potential genetic differences between DVT and PE have not been explored fully in a GWAS. The CTEPH GWAS cohort is likely to comprise more patients that have had a preceding isolated PE than isolated DVT alone.(11) As our CTEPH cohort is enriched for preceding PEs, we may be underpowered to detect more modest genetic associations with healthy controls that would manifest with a DVT enriched cohort. Alternatively, there could be genuine differential genetic associations between PE and CTEPH. A comparative GWAS could be performed to explore this hypothesis. This would require a well characterised PE cohort to define several phenotypic aspects. In addition to defining isolated PE or combined DVT and PE (and recurrence), the radiological resolution of pulmonary arterial perfusion defects should be elucidated to confirm the absence of CTED. Ultimately, access to a deeply phenotyped cohorts of patients with resolved DVT/PE, persistence of CTED and CTEPH would be required to establish pathobiological and genetic differences.

CTEPH is not associated with the genetic determinants of warfarin metabolism. Ineffective anticoagulation of pulmonary embolism due to SNPs associated with warfarin metabolism, is unlikely to be involved in the pathogenesis of CTEPH. Furthermore, genetic differences in the vitamin-K dependent clotting factor pathways that may result in increased thrombus formation prior to the introduction of anticoagulants are also unlikely to be involved in the aetiology of CTEPH.

5.3.3 The effect of genetic *ABO* groups on CTEPH disease severity and survival

Genetic *ABO* groups were not associated with CTEPH disease severity (baseline haemodynamics, 6mwd or WHO functional class) or survival post-PEA. This suggests that the *ABO* association is related to disease aetiology rather than a confounding association from selecting more severe CTEPH phenotypes that would present to specialist pulmonary hypertension recruiting centres. The survival analysis is limited by the relatively small sample size and number of deaths. As peri-operative mortality post-PEA is low in expert centres and long-term mortality is not driven by right ventricular failure (CTEPH), alternative outcome measures need to be considered for future analyses.(7, 15)

5.3.4 CTEPH disease distribution and haemodynamics GWASs

No SNPs were significantly associated in the CTEPH disease subtype (proximal/distal) GWAS. The most significant putative association (rs2313920, OR (95% CI) 0.317 (0.303-0.331), $p=7.63 \times 10^{-8}$) was in *KSR1*. *KSR1* is a gene in chromosome 17 that encodes the protein Kinase suppressor of Ras 1 (KRS1). *KSR1* is expressed in 27 different tissues including heart and lung.(309) KRS1 is an enzyme that functions in multiple receptor tyrosine kinase pathways and participates in the activation of mitogen-activated protein kinase (MAPK) pathways.(310) SNPs in the *KSR1* gene have been associated with monocyte count (rs1105527), inflammatory bowel disease (rs2945412) and other chronic inflammatory diseases (rs10775412).(266, 311, 312) However, there is very little correlation between these SNP associations and the putative *KSR1* SNP associations in the CTEPH GWAS (R^2 : 0.156, 0.153 and 0.042 respectively). The other disease distribution putative association is an intergenic SNP (rs68044424) in chromosome 11 without any known associations in the GWAS catalog or ClinVar database.(140, 313)

There were also no significant associations in the CTEPH haemodynamic (CI, mPAP and PVR) GWASs. The rs4240181 intronic *TUSC3* gene variant in chromosome 8 was the most significant putative association ($p=4.16 \times 10^{-7}$) in the cardiac index GWAS. *TUSC3* is a proposed tumour suppressor gene that encodes the protein Tumour suppressor candidate 3 and is downregulated in some epithelial cancers.(314) *TUSC3* is expressed in 25 different tissues including heart and lung.(309) Rs4240181 has no trait associations in the GWAS Catalog, but SNPs in the *TUSC3* gene have been associated with body mass index and DNA methylation.(315, 316)

The separate GWASs were limited by the sample sizes which would only be adequately powered to detect large effects. This particularly applied to the distal/proximal GWAS, which was limited by the distal disease group size ($n=136$). These analyses should be revisited when the CTEPH GWAS cohort expands and there is adequate power to detect associations.

In summary, there may be differential genetic associations between CTEPH and VTE that could be investigated with a comparative GWAS. *ABO* groups are not associated with CTEPH disease severity or post-PEA survival. Separate GWASs for CTEPH disease subtype and haemodynamics identified putative, but no genome-wide associations.

6 Conclusions

This multi-centre international GWAS is the largest study undertaken in CTEPH and included 1250 CTEPH cases, 1492 healthy controls and ~7 million SNPs. The *ABO* locus was identified as the most significant common variant genetic association in CTEPH in both a discovery and validation cohort. In a joint analysis of both the discovery and validation cohort combined, an intronic variant (rs2519093) in *ABO* was the most significant association with an OR of 2.4 (95% CI 2.3-2.5; $p=3.42 \times 10^{-31}$). Fine mapping using statistical methods identified a 99% credible set of 3 genetic variants that included the SNP (rs507666) that “tags” the A1 genetic *ABO* group. In a subsequent reconstruction of genetic *ABO* groups using haplotypes, the A1 group was enriched in CTEPH patients with the A1A1 group having an odds ratio of 4.4 (95% CI 2.9-6.7) compared with the OO genetic *ABO* group.

ABO is a pleiotropic locus that has been associated with a number of diseases including venous thromboembolism, coronary artery disease and ischaemic stroke.(70, 258) Genetic variation at the *ABO* locus and thrombotic risk has traditionally been attributed to VWF levels, which are 25% lower in O group individuals.(81) Individuals possessing A1 enriched *ABO* groups have higher plasma levels of VWF and factor VIII compared with the O and A2 groups, which is consistent with the enrichment of the A1 group in the current study.(253) However, over 20 plasma protein levels have been associated with genetic variation within the *ABO* locus that relate to immunology, vascular endothelial cell function and coagulation. (272) Therefore, the *ABO* association in the CTEPH GWAS may result in functional effects in addition to mediating VWF levels. Fine mapping of the *ABO* locus using genomic functional annotations revealed the lead SNP is associated with *SURF1* gene expression (eQTL) in the heart. *SURF1* is involved in oxidative phosphorylation, which is another plausible mechanism in CTEPH right ventricular pathobiology.(275) (276)

There was a putative association in the *F11* gene locus (rs2036914) in the discovery cohort however, this was not replicated in the validation cohort. This SNP is also associated with venous thromboembolism in a GWAS, and the absence of a validated

association in the current study may be due to a lack of power.(70) Exploratory gene-based analysis, whereby SNPs are assigned to genes and genome-wide gene association testing is performed, identified a significant association in the *FGG* gene. There was also a putative association ($p < 1 \times 10^{-5}$) signal at the *FGG* locus in the single-variant GWAS analysis. *FGG* is part of the *FGA-FGB-FGG* fibrinogen locus that has been associated with VTE.(70) Therefore, the lack an association in the CTEPH GWAS at the *FGA-FGB-FGG* locus may be due to an under-powered study.

As three quarters of CTEPH patients have had a preceding pulmonary embolism, we may expect similar common variant genetic associations that are seen in VTE.(70) The absence of the VTE genetic associations in CTEPH may be due to a lack of power to detect them (*F11* and *FGA-FGB-FGG*), although the current CTEPH GWAS should be adequately powered to detect the factor V Leiden mutation (rs6025). This may be due to differing genetic associations that predispose to pulmonary embolism or CTEPH and this is discussed further in [Section 7](#). The absence of genetic associations (e.g. factor V Leiden mutation) may also improve understanding of CTEPH pathobiology and rationalise future research. SNPs associated with other types of pulmonary hypertension were not associated with CTEPH, suggesting a lack of shared genetic aetiology that is explored further in [Section 7.1](#). There was no association between genetic determinants of warfarin metabolism and CTEPH suggesting this process is not involved in aetiology or again, reflecting a lack of study power to detect smaller effect sizes.

A large proportion of the variation in VWF levels is genetically determined, with 30% due to *ABO* groups.(200) The *ADAMTS13* gene locus is situated ~200kb downstream of *ABO* and in low-moderate linkage disequilibrium with *ABO*. Whilst the *ADAMTS13* locus was initially associated with CTEPH in the pilot GWAS, this was subsequently determined not to be independent of the *ABO* association.

In a study of 208 patients with CTEPH and 68 health controls, plasma *ADAMTS13* levels were markedly reduced in CTEPH and plasma VWF levels were increased. The *ADAMTS13* reduction in CTEPH was independent of pulmonary hypertension, disease severity or systemic inflammation. These findings implicate dysregulation of the *ADAMTS13*-VWF axis in CTEPH pathobiology. The magnitude of *ADAMTS13*-

VWF dysregulation appears greater in CTEPH than in other thrombotic diseases (e.g. coronary artery disease and ischaemic stroke) suggesting a greater role in CTEPH pathobiology.(193, 212) Furthermore, the combination of low ADAMTS13 and raised VWF has a synergistic effect on the odds of CTEPH. ADAMTS13-VWF dysregulation in CTEPH remains after removal of thromboembolic material during PEA and haemodynamic normalisation suggesting an aetiological role rather than an epiphenomenon. ADAMTS13 is reduced and VWF increased in CTED but there is no marked dysregulation of the axis in pulmonary embolism or idiopathic pulmonary arterial hypertension.

Genetic *ABO* groups had a modest effect on VWF levels in the CTEPH group with some non-O groups having higher VWF however, ADAMTS13 protein levels did not vary by *ABO* group. Given that *ABO* is known to affect VWF with O group individuals having 25% lower levels, this suggests that there are other causes of raised VWF in CTEPH other than differences from genetic *ABO* groups, and conversely suggests *ABO* may be exerting its disease effects via additional mechanisms.(121) By utilising the genomic data, a protein quantitative trait locus was identified near the *ADAMTS13* gene that was associated with ADAMTS13 protein levels and explained ~8% of the variance however, this pQTL is not associated with CTEPH disease risk in the GWAS.

Additional phenotype-genotype analyses did not identify an association between *ABO* genetic groups and CTEPH disease severity or post-PEA survival. This suggests that the *ABO* association is related to disease aetiology rather than a confounding association with disease severity. Separate GWASs for CTEPH disease subtype (proximal *versus* distal chronic thromboembolic distribution) and haemodynamics identified putative, but no genome-wide associations.

The main study limitation was the lack of power in the GWAS to detect genetic variants with lower allele frequencies or more modest effect sizes. This limitation also applied to much of the phenotype-genotype analyses that were predicated on GWAS associations. Whilst the *ABO* locus was associated with CTEPH in discovery and validation cohorts, additional work is required to identify a causal variant and a pathobiological mechanism for its functional effect. The main limitation of the ADAMTS13-VWF work is that whilst dysregulation of the axis was demonstrated in

CTEPH the mechanism by which it is perturbed was not clearly defined. Future studies could address some of these limitations and are discussed further in [Section 7](#).

7 Future research

Areas for future research will be discussed for genome-wide association studies, the ADAMTS13-VWF axis and clinical studies in the following sections.

7.1 GWAS

The CTEPH GWAS had a modest sample size which is a recognised limitation for uncommon diseases. The most important area for future research will be to increase sample size to detect additional common variant associations. Work is currently ongoing to increase CTEPH cases by an additional ~950 and ~4500 for healthy controls. These samples have genotypes available for them from the Affymetrix 6.0 microarray platform. As the current CTEPH GWAS has used a microarray platform using different SNPs, it is difficult to combine them directly for quality control and analysis steps. However, they can be integrated into the current CTEPH GWAS by performing separate GWAS studies, followed by imputation against the same reference platform (to harmonise the SNPs) and then combining datasets prior to statistical association testing. An alternative approach would be to perform completely separate GWAS analyses for cases and controls genotyped on each microarray platform (Illumina and Affymetrix) and then combine the studies using meta-analysis of the summary statistics.⁽¹⁵⁰⁾ The degree to which increasing sample size leads to increasing common variant associations is not clear *a priori* and varies depending on disease.⁽¹²²⁾ A complementary strategy for uncovering additional association signals for uncommon disease is to perform a Bayesian association analysis rather than adopting a traditional frequentist approach, as the strength of evidence with Bayes factors does not vary with sample size or MAF, unlike *p*-values.^(317, 318)

The current study has identified common genetic associations in CTEPH, and previous studies have identified SNPs associated with venous thromboembolism.^(70, 166) However, a key question is what causes the progression from acute PE to CTEPH in a minority of patients. A comparative GWAS study looking at the difference in allele frequencies between resolved PE and CTEPH could uncover further genetic associations leading to mechanistic insight. The ideal comparator group would be PE

patients with radiological evidence of resolution (from the pulmonary vasculature) several months after the acute episode. Current VTE GWASs have used a mixture of predominately deep vein thrombosis and to a less extent pulmonary embolism patients without objective evidence of radiological resolution. Whilst CTEPH is uncommon following PE, there is increasing recognition that post-PE persistence of thrombi and right ventricular impairment occur in up to a third of patients.(9) It would be important to define a precise phenotype for any comparative study with CTEPH and ongoing prospective studies in VTE that include bio-banked blood sample may represent an appropriate cohort.(9, 319) An aim for future studies would be to establish if there is a different genetic risk profile across the spectrum of thromboembolic disease from resolved pulmonary embolism through to post-PE changes, CTED and CTEPH. Cross-trait GWAS could be performed to investigate shared genetic associations with diseases including other forms of pulmonary hypertension (e.g. PAH) and thrombotic diseases (e.g. coronary artery disease and ischaemic stroke) to identify common disease mechanisms.(262)

A limitation of GWAS is the accuracy and breadth of phenotyping which becomes more challenging as sample size increases. An advantage of the current CTEPH GWAS is that the disease has a clear and objective definition that requires multiple radiological modalities and an invasive right heart catheterisation. However, future studies could explore sub-phenotypes further. Whilst the phenotype-genotype analyses in [Chapter 5](#) suggested that genetic associations were not being driven by disease severity, there was still an under-representation of CTEPH patients with less severe disease that did not undergo PEA. The proliferation of electronic health records and ability to perform natural language processing of semi structured radiological reports will be future resources for CTEPH case finding that can be utilised in genomic and other -omic studies.(320) Another potential future resource are large biobanks, such as UK Biobank that have been used for a VTE GWAS however, they would currently lack sufficient CTEPH cases based on published studies of PE in UK Biobank.(166) GWAS using deeper phenotyping may reveal novel genetic associations in disease subtypes such as the distribution of disease in the pulmonary vasculature. Identifying CTEPH subtypes is clinically relevant as subtypes may have different disease mechanisms, natural histories and treatments.(321) Currently, disease distribution is broadly classified as proximal and distal disease, which is of practical importance to guide

stratification for PEA assessment. However, unstandardised discrete categories may not be sufficiently accurate to identify differing disease mechanisms and genetic associations. Improved radiological classification of CTEPH by partitioning the pulmonary vasculature and supported by deep learning would provide more precise phenotypes for future study.(322, 323) CTPA scans also capture data on the lung parenchyma and right heart that are important in CTEPH pathophysiology and would supplement a radiological pulmonary vascular phenotype. Furthermore, radiological phenotypes could be tracked longitudinally from the initial CTPA performed for pulmonary embolism to the CTEPH diagnostic scan in each individual to enrich this novel phenotype. Another strategy could be an analysis to identify unknown CTEPH subtypes using “reverse” GWAS to cluster CTEPH patients using their genomic data with or without additional phenotype data.(324, 325)

Future novel CTEPH-SNP associations that arise from increasing study sample size or additional statistical methods could be interrogated with fine mapping. This can be achieved by increasing SNP density (genetic imputation or resequencing), statistical methods, integration with genomic functional annotations and trans-ethnic fine mapping ([Section 1.5.4](#)). GWAS studies in CTEPH performed across different populations may provide unique insights given observed differences in VTE and CTEPH with ethnicity. VTE is five times more common in black individuals than those with Asian-ancestry, which is not accounted for by known genetic associations.(326) The incidence of CTEPH in various populations are not well described however, there are apparent epidemiological differences between Japanese and Caucasian cohorts.(83) Expanding the CTEPH GWAS to additional ethnic groups would enable trans-ethnic fine mapping and may also reveal novel associations in different ethnicities. An alternative strategy to gain insight into CTEPH pathobiology is to perform a transcriptome wide association study (TWAS). This involves leveraging genotypes from a GWAS and combining them with gene expression datasets to identify gene-trait associations.(327, 328) The advantage of this approach is that it does not require directly profiling gene expressions across multiple tissues in all GWAS individuals.(327)

Genotypes from GWAS can be used in additional analyses to investigate causal factors in CTEPH. Inflammation has been associated with CTEPH, but it is unclear if

this is causal or epiphenomenon. Plasma C-reactive protein is raised in CTEPH and has been implicated in disease pathobiology.(50, 51) Mendelian randomisation is a method of investigating causal relationships using GWAS risk variants as genetic instruments ([Section 1.5.1](#)). Data on CRP could be collected for the CTEPH GWAS cohort combined with CRP data from a healthy control group (i.e. UK Biobank) and Mendelian randomisation performed to explore the potential causal relationship between CRP and CTEPH. This strategy has been used to investigate the role of CRP in coronary artery disease and more recently, the role of red cell distribution width in pulmonary arterial hypertension.(139, 329)

7.2 ADAMTS13-VWF

The ADAMTS13-VWF axis was demonstrated to be dysregulated in CTEPH but the mechanisms by which ADAMTS13 was decreased and VWF increased were not fully elucidated. In addition to quantifying ADAMTS13 in vascular endothelial cells in selected tissues that was described [Section 4.3.6](#), the site of ADAMTS13 reduction could be further explored. ADAMTS13 is predominately produced by hepatic stellate cells in the liver with a contribution from vascular endothelial cells.(177) To explore whether ADAMTS13 production is reduced by either tissue, reprogrammed induced pluripotent stem cells (iPSC) could be utilised.(330) Blood samples from CTEPH patients could be used to generate iPSCs and then derived hepatic stellate cells and vascular endothelial cells produced.(331, 332) The advantage of this approach is that the derived cell lines would have the same genotype as the CTEPH individual from whom the cells were derived. This would enable an exploration of the effect of common genetic variants on ADAMTS13 production in the liver and vascular endothelium with the ability to stratify experiments by plasma ADAMTS13-VWF levels and genotypes. If no change in the production of ADAMTS13 is demonstrated in CTEPH patients compared with control cells, then the reduction of ADAMTS13 may be related to consumption, sequestration or excretion and these areas could be investigated further.

Experimental studies of CTEPH using small animal models have not fully recapitulated the chronic changes that occur in the pulmonary vasculature or right heart.(55) This is partly due to differential vascular responses to chronic pulmonary vascular

obstruction across species.(55) One of the most promising animal models is a porcine model that could be used for future mechanistic work.(333) Ideally a CTEPH animal model could be developed with either ADAMTS13 and/or VWF knockouts or with an induced alteration in ADAMTS13-VWF levels. Whilst *Adamts13^{-/-}* and *Vwf^{-/-}* knockouts exist for the study of stroke and TTP in mice, there are currently no robust large animal models.(334) Stroke murine models with perturbed ADAMT13-VWF have revealed increased inflammatory cells and increased infarct size in the brains of *Adamts13^{-/-}* mice that are improved by an infusion of recombinant human ADAMTS13 (rhADAMTS13).(290, 335) Future studies in CTEPH could explore the mechanisms of immunothrombosis using similar methodology in an appropriately developed animal model.

Whilst *ADAMTS13* common variants were not associated with CTEPH in the GWAS, this methodology would not have detected rare variant associations. Rare *ADAMTS13* variants were overrepresented in a small study of VTE patients and future studies could explore if they are overrepresented in CTEPH compared with healthy controls and VTE.(298) Exome sequencing of the *ADAMTS13* gene could be performed in CTEPH patients enriched with the lowest ADAMTS13 plasma levels.

7.3 Clinical perspectives

Clinical prediction scores for CTEPH following acute PE do not currently incorporate blood biomarkers. Future studies using robustly phenotyped PE cohorts to ascertain the presence and extent of residual perfusion defects, could investigate if the ADAMTS13-VWF axis varies in the spectrum of disease encompassing post-PE syndrome. Determining if ADAMTS13-VWF axis dysregulation precedes the development of chronic thromboembolic pathology could inform CTEPH risk stratification. In population studies, healthy individuals with low ADAMTS13 have increased risk of developing ischaemic stroke, which provides a rationale for prospective studies following PE.(336) However, low ADAMTS13 alone is unlikely to be sufficiently robust to predict CTEPH following acute PE, given that even patients with severe ADAMTS13 deficiency do not always develop TTP and additional environmental and/or genetic contributors are required.(337) A risk prediction score utilising clinical variables has been developed in CTEPH but is yet to be

validated.(172) Whether this CTEPH risk prediction model could be improved with the addition of ADAMTS13-VWF levels and genetic *ABO* group together with established clinical risk factors could be investigated in subsequent studies. If ADAMTS13-VWF levels post-PE enable patients to be stratified into risk groups then this could inform future clinical drug trials. Currently, pulmonary embolism is treated with anticoagulation to prevent further VTE, and if there is haemodynamic compromise, thrombolytic therapies that break down blood clots are considered.(198) Thrombolysis improves right ventricular function in acute PE but does not prevent CTEPH and is associated with bleeding complications.(18) Therefore, alternative treatment strategies will require future evaluation including whether ADAMTS13-VWF levels following acute PE can be used to stratify CTEPH preventative treatment modalities and durations. In stroke murine models, rhADAMTS13 decreases infarct size and is not associated with the excess major bleeding seen with tissue plasminogen activator (tPA) thrombolysis.(290) Furthermore, some thrombotic occlusions in stroke are resistant to thrombolysis with tPA but can be thrombolysed with ADAMTS13 infusions.(338) Recombinant ADAMTS13 improved neovascularisation, vascular repair and stroke outcome in murine models even when administered in the recovery phase at 7 days.(297) Moreover, in stroke patients plasma ADAMTS13 can predict response to thrombolytic reperfusion strategies.(339) Therefore, if risk prediction for developing CTEPH following acute PE improves, then future studies could address the effect of rhADAMTS13 infusions in a high risk stratified PE population to prevent CTEPH.

References

1. Galie N, Humbert M, Vachiery JL, Gibbs S, Lang I, Torbicki A, et al. 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). *Eur Heart J*. 2016;37(1):67-119.
2. Matthews DT, Hemnes AR. Current concepts in the pathogenesis of chronic thromboembolic pulmonary hypertension. *Pulm Circ*. 2016;6(2):145-54.
3. Moser KM, Bloor CM. Pulmonary vascular lesions occurring in patients with chronic major vessel thromboembolic pulmonary hypertension. *Chest*. 1993;103(3):685-92.
4. Galie N, Kim HS. Pulmonary microvascular disease in chronic thromboembolic pulmonary hypertension. *Proc Am Thorac Soc*. 2006;3(7):571-6.
5. Naess IA, Christiansen SC, Romundstad P, Cannegieter SC, Rosendaal FR, Hammerstrom J. Incidence and mortality of venous thrombosis: a population-based study. *J Thromb Haemost*. 2007;5(4):692-9.
6. Ende-Verhaar YM, Cannegieter SC, Vonk Noordegraaf A, Delcroix M, Pruszczyk P, Mairuhu AT, et al. Incidence of chronic thromboembolic pulmonary hypertension after acute pulmonary embolism: a contemporary view of the published literature. *Eur Respir J*. 2017;49(2):1601792.
7. Jenkins D, Madani M, Fadel E, D'Armini AM, Mayer E. Pulmonary endarterectomy in the management of chronic thromboembolic pulmonary hypertension. *Eur Respir Rev*. 2017;26(143):160111.
8. Gopalan D, Delcroix M, Held M. Diagnosis of chronic thromboembolic pulmonary hypertension. *Euro Respir Rev*. 2017;26(143):160108.
9. Klok FA, van der Hulle T, den Exter PL, Lankeit M, Huisman MV, Konstantinides S. The post-PE syndrome: a new concept for chronic complications of pulmonary embolism. *Blood Rev*. 2014;28(6):221-6.
10. Taboada D, Pepke-Zaba J, Jenkins DP, Berman M, Treacy CM, Cannon JE, et al. Outcome of pulmonary endarterectomy in symptomatic chronic thromboembolic disease. *Eur Respir J*. 2014;44(6):1635-45.
11. Pepke-Zaba J, Delcroix M, Lang I, Mayer E, Jansa P, Ambroz D, et al. Chronic thromboembolic pulmonary hypertension (CTEPH): results from an international prospective registry. *Circulation*. 2011;124(18):1973-81.
12. Bonderman D, Wilkens H, Wakounig S, Schafers HJ, Jansa P, Lindner J, et al. Risk factors for chronic thromboembolic pulmonary hypertension. *Eur Respir J*. 2009;33(2):325-31.
13. Zhang M, Wang N, Zhai Z, Zhou R, Liu Y, Yang Y. Incidence and risk factors of chronic thromboembolic pulmonary hypertension after acute pulmonary embolism: a systematic review and meta-analysis of cohort studies. *J Thorac Dis*. 2018;10(8):4751-63.
14. Lang IM, Simonneau G, Pepke-Zaba JW, Mayer E, Ambroz D, Blanco I, et al. Factors associated with diagnosis and operability of chronic thromboembolic pulmonary hypertension. A case-control study. *Thromb Haemost*. 2013;110(1):83-91.

15. Cannon JE, Su L, Kiely DG, Page K, Toshner M, Swietlik E, et al. Dynamic Risk Stratification of Patient Long-Term Outcome After Pulmonary Endarterectomy: Results From the United Kingdom National Cohort. *Circulation*. 2016;133(18):1761-71.
16. Bunclark K, Newnham M, Chiu Y-D, Ruggiero A, Villar SS, Cannon JE, et al. A multicenter study of anticoagulation in operable chronic thromboembolic pulmonary hypertension. *J Thromb Haemost*. 2020;18(1):114-22.
17. Kim NH, Delcroix M, Jais X, Madani MM, Matsubara H, Mayer E, et al. Chronic thromboembolic pulmonary hypertension. *Eur Respir J*. 2019;53(1):1801915.
18. Konstantinides SV, Vicaut E, Danays T, Becattini C, Bertolotti L, Beyer-Westendorf J, et al. Impact of Thrombolytic Therapy on the Long-Term Outcome of Intermediate-Risk Pulmonary Embolism. *J Am Coll Cardiol*. 2017;69(12):1536-44.
19. Jenkins D. Pulmonary endarterectomy: the potentially curative treatment for patients with chronic thromboembolic pulmonary hypertension. *Euro Respir Rev*. 2015;24(136):263-71.
20. Newnham M, Hernandez-Sanchez J, Dunning J, Ng C, Tsui S, Bunclark K, et al. Age should not be a barrier for pulmonary endarterectomy in carefully selected patients. *Eur Respir J*. 2017;50(6):1701804.
21. Shimura N, Kataoka M, Inami T, Yanagisawa R, Ishiguro H, Kawakami T, et al. Additional percutaneous transluminal pulmonary angioplasty for residual or recurrent pulmonary hypertension after pulmonary endarterectomy. *Int J Cardiol*. 2015;183:138-42.
22. Robbins IM, Pugh ME, Hemnes AR. Update on chronic thromboembolic pulmonary hypertension. *Trends Cardiovasc Med*. 2017;27(1):29-37.
23. Tanabe N, Kawakami T, Satoh T, Matsubara H, Nakanishi N, Ogino H, et al. Balloon pulmonary angioplasty for chronic thromboembolic pulmonary hypertension: A systematic review. *Respir Investig*. 2018;56(4):332-41.
24. Ghofrani HA, D'Armini AM, Grimminger F, Hoeper MM, Jansa P, Kim NH, et al. Riociguat for the treatment of chronic thromboembolic pulmonary hypertension. *N Engl J Med*. 2013;369(4):319-29.
25. Ghofrani HA, Simonneau G, D'Armini AM, Fedullo P, Howard LS, Jais X, et al. Macitentan for the treatment of inoperable chronic thromboembolic pulmonary hypertension (MERIT-1): results from the multicentre, phase 2, randomised, double-blind, placebo-controlled study. *The Lancet Respir Med*. 2017;5(10):785-94.
26. Madani MM, Jamieson SW. Technical advances of pulmonary endarterectomy for chronic thromboembolic pulmonary hypertension. *Semin Thorac Cardiovasc Surg*. 2006;18(3):243-9.
27. Simonneau G, Torbicki A, Dorfmueller P, Kim N. The pathophysiology of chronic thromboembolic pulmonary hypertension. *Eur Respir Rev*. 2017;26(143):160112.
28. Bernard J, Yi ES. Pulmonary thromboendarterectomy: a clinicopathologic study of 200 consecutive pulmonary thromboendarterectomy cases in one institution. *Hum Pathol*. 2007;38(6):871-7.
29. Blauwet LA, Edwards WD, Tazelaar HD, McGregor CG. Surgical pathology of pulmonary thromboendarterectomy: a study of 54 cases from 1990 to 2001. *Hum Pathol*. 2003;34(12):1290-8.
30. St Croix CM, Steinhorn RH. New Thoughts about the Origin of Plexiform Lesions. *Am J Respir Crit Care Med*. 2016;193(5):484-5.

31. Dorfmueller P, Gunther S, Ghigna MR, Thomas de Montpreville V, Boulate D, Paul JF, et al. Microvascular disease in chronic thromboembolic pulmonary hypertension: a role for pulmonary veins and systemic vasculature. *Eur Respir J*. 2014;44(5):1275-88.
32. Lang IM, Dorfmueller P, Vonk Noordegraaf A. The Pathobiology of Chronic Thromboembolic Pulmonary Hypertension. *Ann Am Thorac Soc*. 2016;13 Suppl 3:S215-21.
33. Gale AJ. Continuing education course #2: current understanding of hemostasis. *Toxicol Pathol*. 2011;39(1):273-80.
34. Wolf M, Boyer-Neumann C, Parent F, Eschwege V, Jaillet H, Meyer D, et al. Thrombotic risk factors in pulmonary hypertension. *Eur Respir J*. 2000;15(2):395-9.
35. Bonderman D, Turecek PL, Jakowitsch J, Weltermann A, Adlbrecht C, Schneider B, et al. High prevalence of elevated clotting factor VIII in chronic thromboembolic pulmonary hypertension. *Thromb and haemost*. 2003;90(3):372-6.
36. Olman MA, Marsh JJ, Lang IM, Moser KM, Binder BR, Schleef RR. Endogenous fibrinolytic system in chronic large-vessel thromboembolic pulmonary hypertension. *Circulation*. 1992;86(4):1241-8.
37. Wong L, Szydlo R, Gibbs S, Laffan M. Hereditary and acquired thrombotic risk factors for chronic thromboembolic pulmonary hypertension. *Blood Coagul Fibrinolysis*. 2010;21(3):201-6
38. Wikipedia. 2020. Coagulation. [Online]. [Accessed 1 March 2020]. Available from: <https://en.wikipedia.org/wiki/Coagulation>
39. Morris TA, Marsh JJ, Chiles PG, Auger WR, Fedullo PF, Woods VL, Jr. Fibrin derived from patients with chronic thromboembolic pulmonary hypertension is resistant to lysis. *Am J Respir Crit Care Med*. 2006;173(11):1270-5.
40. Morris A, Marsh J, Chiles G, Magaña M, Liang N-C, Soler X, et al. High prevalence of dysfibrinogenemia among patients with chronic thromboembolic pulmonary hypertension. *Blood*. 2009;114(9):1929-36.
41. Miniati M, Fiorillo C, Becatti M, Monti S, Bottai M, Marini C, et al. Fibrin resistance to lysis in patients with pulmonary hypertension other than thromboembolic. *Am J Respir Crit Care Med*. 2010;181(9):992-6.
42. Li JF, Lin Y, Yang YH, Gan HL, Liang Y, Liu J, et al. Fibrinogen A α Thr312Ala polymorphism specifically contributes to chronic thromboembolic pulmonary hypertension by increasing fibrin resistance. *PLoS One*. 2013;8(7):e69635.
43. Yaoita N, Satoh K, Satoh T, Sugimura K, Tatebe S, Yamamoto S, et al. Thrombin-Activatable Fibrinolysis Inhibitor in Chronic Thromboembolic Pulmonary Hypertension. *Arterioscler Thromb Vasc Biol*. 2016;36(6):1293-301.
44. Wikipedia. 2020. Fibrinolysis. [Online]. [Accessed 1 March 2020]. Available from: <https://en.wikipedia.org/wiki/Fibrinolysis>
45. Alias S, Redwan B, Panzenboeck A, Winter MP, Schubert U, Voswinckel R, et al. Defective angiogenesis delays thrombus resolution: a potential pathogenetic mechanism underlying chronic thromboembolic pulmonary hypertension. *Arterioscler Thromb Vasc Biol*. 2014;34(4):810-9.
46. Quarck R, Wynants M, Verbeken E, Meyns B, Delcroix M. Contribution of inflammation and impaired angiogenesis to the pathobiology of chronic thromboembolic pulmonary hypertension. *Eur Respir J*. 2015;46(2):431-43.
47. Zabini D, Nagaraj C, Stacher E, Lang IM, Nierlich P, Klepetko W, et al. Angiostatic factors in the pulmonary endarterectomy material from chronic thromboembolic pulmonary hypertension patients cause endothelial dysfunction. *PLoS One*. 2012;7(8):e43793.

48. Price LC, Wort SJ, Perros F, Dorfmüller P, Huertas A, Montani D, et al. Inflammation in pulmonary arterial hypertension. *Chest*. 2012;141(1):210-21.
49. Bonderman D, Jakowitsch J, Redwan B, Bergmeister H, Renner MK, Panzenbock H, et al. Role for staphylococci in misguided thrombus resolution of chronic thromboembolic pulmonary hypertension. *Arterioscler Thromb Vasc Biol*. 2008;28(4):678-84.
50. Quarck R, Nawrot T, Meyns B, Delcroix M. C-reactive protein: a new predictor of adverse outcome in pulmonary arterial hypertension. *J Am Coll Cardiol*. 2009;53(14):1211-8.
51. Wynants M, Quarck R, Ronisz A, Alfaro-Moreno E, Van Raemdonck D, Meyns B, et al. Effects of C-reactive protein on human pulmonary vascular cells in chronic thromboembolic pulmonary hypertension. *Eur Respir J*. 2012;40(4):886-94.
52. Kimura H, Okada O, Tanabe N, Tanaka Y, Terai M, Takiguchi Y, et al. Plasma monocyte chemoattractant protein-1 and pulmonary vascular resistance in chronic thromboembolic pulmonary hypertension. *Am J Respir Crit Care Med*. 2001;164(2):319-24.
53. Zabini D, Heinemann A, Foris V, Nagaraj C, Nierlich P, Balint Z, et al. Comprehensive analysis of inflammatory markers in chronic thromboembolic pulmonary hypertension patients. *Eur Respir J*. 2014;44(4):951-62.
54. Toshner M, Pepke-Zaba J. Chronic thromboembolic pulmonary hypertension: time for research in pathophysiology to catch up with developments in treatment. *F1000Prime Rep*. 2014;6:38.
55. Mercier O, Fadel E. Chronic thromboembolic pulmonary hypertension: animal models. *Eur Respir J*. 2013;41(5):1200-6.
56. MacCallum P, Bowles L, Keeling D. Diagnosis and management of heritable thrombophilias. *BMJ*. 2014;349:g4387.
57. Hotoleanu C. Genetic Risk Factors in Venous Thromboembolism. *Adv Exp Med Biol*. 2017;906:253-72.
58. Khan S, Dickerman JD. Hereditary thrombophilia. *Thromb J*. 2006;4:15.
59. Bloomenthal D, von Dadelszen P, Liston R, Magee L, Tsang P. The effect of factor V Leiden carriage on maternal and fetal health. *CMAJ*. 2002;167(1):48-54.
60. Jadaon MM. Epidemiology of Prothrombin G20210A Mutation in the Mediterranean Region. *Mediterr J Hematol Infect Dis*. 2011;3(1):e2011054.
61. Khan MM, Motto DG, Lentz SR, Chauhan AK. ADAMTS13 reduces VWF-mediated acute inflammation following focal cerebral ischemia in mice. *J thromb haemost*. 2012;10(8):1665-71.
62. Dinarvand P, Moser KA. Protein C Deficiency. *Arch Pathol Lab Med*. 2019;143(10):1281-5.
63. ten Kate MK, van der Meer J. Protein S deficiency: a clinical perspective. *Haemophilia*. 2008;14(6):1222-8.
64. Patnaik MM, Moll S. Inherited antithrombin deficiency: a review. *Haemophilia*. 2008;14(6):1229-39.
65. Lang IM, Klepetko W, Pabinger I. No increased prevalence of the factor V Leiden mutation in chronic major vessel thromboembolic pulmonary hypertension (CTEPH). *Thromb Haemost*. 1996;76(3):476-7.
66. Suntharalingam J, Goldsmith K, van Marion V, Long L, Treacy CM, Dudbridge F, et al. Fibrinogen Aα Thr312Ala polymorphism is associated with chronic thromboembolic pulmonary hypertension. *Eur Respir J*. 2008;31(4):736-41.
67. Franchini M, Makris M. Non-O blood group: an important genetic risk factor for venous thromboembolism. *Blood Transfus*. 2013;11(2):164-5.

68. Delcroix M, Lang I, Pepke-Zaba J, Jansa P, D'Armini AM, Snijder R, et al. Long-Term Outcome of Patients With Chronic Thromboembolic Pulmonary Hypertension: Results From an International Prospective Registry. *Circulation*. 2016;133(9):859-71.
69. Lotta LA, Wang M, Yu J, Martinelli I, Yu F, Passamonti SM, et al. Identification of genetic risk variants for deep vein thrombosis by multiplexed next-generation sequencing of 186 hemostatic/pro-inflammatory genes. *BMC Med Genomics*. 2012;5:7.
70. Germain M, Chasman DI, de Haan H, Tang W, Lindstrom S, Weng LC, et al. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. *Am J Hum Genet*. 2015;96(4):532-42.
71. Rasmussen-Torvik LJ, Cushman M, Tsai MY, Zhang Y, Heckbert SR, Rosamond WD, et al. The association of alpha-fibrinogen Thr312Ala polymorphism and venous thromboembolism in the LITE study. *Thromb Res*. 2007;121(1):1-7.
72. Humbert M, Guignabert C, Bonnet S, Dorfmueller P, Klinger JR, Nicolls MR, et al. Pathology and pathobiology of pulmonary hypertension: state of the art and research perspectives. *Eur Respir J*. 2019;53(1):1801887.
73. Simonneau G, Montani D, Celermajer DS, Denton CP, Gatzoulis MA, Krowka M, et al. Haemodynamic definitions and updated clinical classification of pulmonary hypertension. *Eur Respir J*. 2019;53(1): 1801913.
74. Thenappan T, Ormiston ML, Ryan JJ, Archer SL. Pulmonary arterial hypertension: pathogenesis and clinical management. *BMJ*. 2018;360:j5492.
75. Southgate L, Machado RD, Gräf S, Morrell NW. Molecular genetic framework underlying pulmonary arterial hypertension. *Nat Rev Cardiol*. 2020;17(2):85-95.
76. Morrell NW, Aldred MA, Chung WK, Elliott CG, Nichols WC, Soubrier F, et al. Genetics and genomics of pulmonary arterial hypertension. *Eur Respir J*. 2019;53(1): 1801899.
77. Xi Q, Liu Z, Zhao Z, Luo Q, Huang Z. High Frequency of Pulmonary Hypertension-Causing Gene Mutation in Chinese Patients with Chronic Thromboembolic Pulmonary Hypertension. *PLoS One*. 2016;11(1):e0147396.
78. Suntharalingam J, Machado RD, Sharples LD, Toshner MR, Sheares KK, Hughes RJ, et al. Demographic features, BMPR2 status and outcomes in distal chronic thromboembolic pulmonary hypertension. *Thorax*. 2007;62(7):617-22.
79. Ulrich S, Szamalek-Hoegel J, Hersberger M, Fischler M, Garcia JS, Huber LC, et al. Sequence variants in BMPR2 and genes involved in the serotonin and nitric oxide pathways in idiopathic pulmonary arterial hypertension and chronic thromboembolic pulmonary hypertension: relation to clinical parameters and comparison with left heart disease. *Respiration*. 2010;79(4):279-87.
80. Stearman RS, Cornelius AR, Lu X, Conklin DS, Del Rosario MJ, Lowe AM, et al. Functional prostacyclin synthase promoter polymorphisms. Impact in pulmonary arterial hypertension. *Am J Respir Crit Care Med*. 2014;189(9):1110-20.
81. Amano S, Tatsumi K, Tanabe N, Kasahara Y, Kurosu K, Takiguchi Y, et al. Polymorphism of the promoter region of prostacyclin synthase gene in chronic thromboembolic pulmonary hypertension. *Respirology*. 2004;9(2):184-9.
82. Graf S, Haimel M, Bleda M, Hadinnapola C, Southgate L, Li W, et al. Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. *Nat Commun*. 2018;9(1):1416.
83. Nakamura M, Okada O, Sakuma M, Nakanishi N, Miyahara Y, Yamada N, et al. Incidence and clinical characteristics of chronic pulmonary thromboembolism in

- Japan compared with acute pulmonary thromboembolism: results of a multicenter registry of the Japanese Society of Pulmonary Embolism Research. *Circ J*. 2002;66(3):257-60.
84. Kominami S, Tanabe N, Ota M, Naruse K, Katsuyama Y, Nakanishi N, et al. HLA-DPB1 and NFKBIL1 may confer the susceptibility to chronic thromboembolic pulmonary hypertension in the absence of deep vein thrombosis. *J Hum Genet*. 2009;54(2):108-14.
 85. Tanabe N, Kimura A, Amano S, Okada O, Kasahara Y, Tatsumi K, et al. Association of clinical features with HLA in chronic pulmonary thromboembolism. *Eur Respir J*. 2005;25(1):131-8.
 86. Reinders J, Rozemuller EH, van Gent R, Arts-Hilkes YH, van den Tweel JG, Tilanus MG. Extended HLA-DPB1 polymorphism: an RNA approach for HLA-DPB1 typing. *Immunogenetics*. 2005;57(10):790-4.
 87. Albertella MR, Campbell RD. Characterization of a novel gene in the human major histocompatibility complex that encodes a potential new member of the I kappa B family of proteins. *Hum Mol Genet*. 1994;3(5):793-9.
 88. Tanabe N, Amano S, Tatsumi K, Kominami S, Igarashi N, Shimura R, et al. Angiotensin-converting enzyme gene polymorphisms and prognosis in chronic thromboembolic pulmonary hypertension. *Circ J*. 2006;70(9):1174-9.
 89. Zou L, Li W, Han J, Yang Y, Jin J, Xiao F, et al. Identification of a low frequency missense mutation in MUC6 contributing to pulmonary artery hypertension by whole-exome sequencing. *Pulm Circ*. 2018;8(3):2045894018794374.
 90. Desmarais J, Elliott CG. Familial Chronic Thromboembolic Pulmonary Hypertension. *Chest*. 2016;149(4):e99-e101.
 91. Kataoka M, Momose Y, Aimi Y, Fukuda K, Gamou S, Satoh T. Familial Chronic Thromboembolic Pulmonary Hypertension in a Pair of Japanese Brothers. *Chest*. 2016;150(3):748-9.
 92. Dodson MW, Allen-Brady K, Brown LM, Elliott CG, Cannon-Albright LA. Chronic Thromboembolic Pulmonary Hypertension Cases Cluster in Families. *Chest*. 2019;155(2):384-90.
 93. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. *PLoS computational biology*. 2012;8(12):e1002822.
 94. Rhodes CJ, Batai K, Bleda M, Haimel M, Southgate L, Germain M, et al. Genetic determinants of risk in pulmonary arterial hypertension: international genome-wide association studies and meta-analysis. *Lancet Respir Med*. 2019;7(3):227-38.
 95. Opitz I, Kirschner MB. Molecular Research in Chronic Thromboembolic Pulmonary Hypertension. *Int J Mol Sci*. 2019;20(3):784.
 96. Gu S, Su P, Yan J, Zhang X, An X, Gao J, et al. Comparison of gene expression profiles and related pathways in chronic thromboembolic pulmonary hypertension. *Int J Mol Med*. 2014;33(2):277-300.
 97. Guo L, Yang Y, Liu J, Wang L, Li J, Wang Y, et al. Differentially expressed plasma microRNAs and the potential regulatory function of Let-7b in chronic thromboembolic pulmonary hypertension. *PloS one*. 2014;9(6): e101055.
 98. Wang Y, Huang X, Leng D, Li J, Wang L, Liang Y, et al. DNA methylation signatures of pulmonary arterial smooth muscle cells in chronic thromboembolic pulmonary hypertension. *Physiol Genomics*. 2018;50(5):313-22.
 99. Deng C, Wu D, Yang M, Chen Y, Wang C, Zhong Z, et al. Expression of tissue factor and forkhead box transcription factor O-1 in a rat model for chronic thromboembolic pulmonary hypertension. *J Thromb Thrombolysis*. 2016;42(4):520-8.

100. Crous-Bou M, Harrington LB, Kabrhel C. Environmental and Genetic Risk Factors Associated with Venous Thromboembolism. *Semin Thromb Hemost.* 2016;42(8):808-20.
101. Heit JA, Phelps MA, Ward SA, Slusser JP, Petterson TM, De Andrade M. Familial segregation of venous thromboembolism. *J Thromb Haemost.* 2004;2(5):731-6.
102. Margaglione M, Grandone E. Population genetics of venous thromboembolism. A narrative review. *Thromb Haemost.* 2011;105(2):221-31.
103. Rosendaal FR, Reitsma PH. Genetics of venous thrombosis. *J Thromb Haemost.* 2009;7 Suppl 1:301-4.
104. Gohil R, Peck G, Sharma P. The genetics of venous thromboembolism. A meta-analysis involving approximately 120,000 cases and 180,000 controls. *Thromb Haemost.* 2009;102(2):360-70.
105. Uitte de Willige S, de Visser MC, Houwing-Duistermaat JJ, Rosendaal FR, Vos HL, Bertina RM. Genetic variation in the fibrinogen gamma gene increases the risk for deep venous thrombosis by reducing plasma fibrinogen gamma' levels. *Blood.* 2005;106(13):4176-83.
106. Emmerich J, Rosendaal FR, Cattaneo M, Margaglione M, De Stefano V, Cumming T, et al. Combined effect of factor V Leiden and prothrombin 20210A on the risk of venous thromboembolism--pooled analysis of 8 case-control studies including 2310 cases and 3204 controls. Study Group for Pooled-Analysis in Venous Thromboembolism. *Thromb Haemost.* 2001;86(3):809-16.
107. Gerhardt A, Scharf RE, Beckmann MW, Struve S, Bender HG, Pillny M, et al. Prothrombin and factor V mutations in women with a history of thrombosis during pregnancy and the puerperium. *N Engl J Med.* 2000;342(6):374-80.
108. Vandenbroucke JP, Koster T, Briët E, Reitsma PH, Bertina RM, Rosendaal FR. Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. *Lancet.* 1994;344(8935):1453-7.
109. Cushman M, Kuller LH, Prentice R, Rodabough RJ, Psaty BM, Stafford RS, et al. Estrogen plus progestin and risk of venous thrombosis. *JAMA.* 2004;292(13):1573-80.
110. Stevens SM, Woller SC, Bauer KA, Kasthuri R, Cushman M, Streiff M, et al. Guidance for the evaluation and treatment of hereditary and acquired thrombophilia. *J Thromb Thrombolysis.* 2016;41(1):154-64.
111. Kearon C. Natural history of venous thromboembolism. *Circulation.* 2003;107(23 Suppl 1):I22-30.
112. van Stralen KJ, Doggen CJ, Bezemer ID, Pomp ER, Lisman T, Rosendaal FR. Mechanisms of the factor V Leiden paradox. *Arterioscler Thromb Vasc Biol.* 2008;28(10):1872-7.
113. Dentali F, Ageno W, Bozzato S, Malato A, Gianni M, Squizzato A, et al. Role of factor V Leiden or G20210A prothrombin mutation in patients with symptomatic pulmonary embolism and deep vein thrombosis: a meta-analysis of the literature. *J Thromb Haemost.* 2012;10(4):732-7.
114. Dentali F, Sironi AP, Ageno W, Turato S, Bonfanti C, Frattini F, et al. Non-O blood type is the commonest genetic risk factor for VTE: results from a meta-analysis of the literature. *Semin Thromb Hemost.* 2012;38(5):535-48.
115. Ohira T, Cushman M, Tsai MY, Zhang Y, Heckbert SR, Zakai NA, et al. ABO blood group, other risk factors and incidence of venous thromboembolism: the Longitudinal Investigation of Thromboembolism Etiology (LITE). *J Thromb Haemost.* 2007;5(7):1455-61.

116. Gandara E, Kovacs MJ, Kahn SR, Wells PS, Anderson DA, Chagnon I, et al. Non-OO blood type influences the risk of recurrent venous thromboembolism. A cohort study. *Thromb Haemost.* 2013;110(6):1172-9.
117. Wolpin BM, Kabrhel C, Varraso R, Kraft P, Rimm EB, Goldhaber SZ, et al. Prospective study of ABO blood type and the risk of pulmonary embolism in two large cohort studies. *Thromb Haemost.* 2010;104(5):962-71.
118. Jenkins PV, O'Donnell JS. ABO blood group determines plasma von Willebrand factor levels: a biologic function after all? *Transfusion.* 2006;46(10):1836-44.
119. Seltsam A, Hallensleben M, Kollmann A, Blasczyk R. The nature of diversity and diversification at the ABO locus. *Blood.* 2003;102(8):3035-42.
120. O'Donnell J, Laffan MA. The relationship between ABO histo-blood group, factor VIII and von Willebrand factor. *Transfus Med.* 2001;11(4):343-51.
121. Gill JC, Endres-Brooks J, Bauer PJ, Marks WJ, Jr., Montgomery RR. The effect of ABO blood group on the diagnosis of von Willebrand disease. *Blood.* 1987;69(6):1691-5.
122. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet.* 2017;101(1):5-22.
123. Manolio TA. Genomewide association studies and assessment of the risk of disease. *New Engl J Med.* 2010;363(2):166-76.
124. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet.* 2012;90(1):7-24.
125. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. *Nat Med.* 2015;526(7571):68-74.
126. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* 2017;18(1):77.
127. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet.* 2001;17(9):502-10.
128. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. *Science.* 2005;308(5720):385-9.
129. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids Res.* 2019;47(D1):D1005-D12.
130. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum Mol Genet.* 2018;27(20):3641-9.
131. Ioannidis JP, Tarone R, McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology.* 2011;22(4):450-6.
132. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA.* 2008;299(11):1335-44.
133. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nat Med.* 2018;562(7726):203-9.
134. Gudbjartsson DF, Arnar DO, Helgadottir A, Gretarsdottir S, Holm H, Sigurdsson A, et al. Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nat Med.* 2007;448(7151):353-7.

135. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet.* 2018;19(9):581-90.
136. Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat Rev Genet.* 2016;17(3):129-45.
137. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. *Nat Med.* 2020;577(7789):179-89.
138. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ.* 2018;362:k601.
139. Wensley F, Gao P, Burgess S, Kaptoge S, Di Angelantonio E, Shah T, et al. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ.* 2011;342:d548.
140. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 2014;42:D1001-6.
141. LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.* 2009;37(13):4181-93.
142. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc.* 2010;5(9):1564-73.
143. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev Genet.* 2008;9(5):356-69.
144. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet.* 2008;9(6):477-85.
145. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. Basic statistical analysis in genetic case-control studies. *Nat Protoc.* 2011;6(2):121-33.
146. Pettersson FH, Anderson CA, Clarke GM, Barrett JC, Cardon LR, Morris AP, et al. Marker selection for genetic case-control association studies. *Nat Protoc.* 2009;4(5):743-52.
147. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet.* 2003;361(9357):598-604.
148. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet.* 2004;36(5):512-7.
149. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet.* 2010;11(7):499-511.
150. Zeggini E, Ioannidis JP. Meta-analysis in genome-wide association studies. *Pharmacogenomics.* 2009;10(2):191-201.
151. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nat Med.* 2009;461(7265):747-53.
152. Visscher PM, Medland SE, Ferreira MA, Morley KI, Zhu G, Cornes BK, et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* 2006;2(3):e41.
153. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell.* 2017;169(7):1177-86.

154. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* 2010;42(7):565-9.
155. Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. Rare and low-frequency coding variants alter human adult height. *Nat Med.* 2017;542(7640):186-90.
156. Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Hum Mol Genet.* 2015;24(R1):R102-10.
157. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet.* 2015;47(11):1236-41.
158. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet.* 2018;19(8):491-504.
159. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Methods in health service research. An introduction to bayesian methods in health technology assessment. *BMJ.* 1999;319(7208):508-12.
160. Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet.* 2012;44(12):1294-301.
161. Wang X, Liu X, Sim X, Xu H, Khor CC, Ong RT, et al. A statistical method for region-based meta-analysis of genome-wide association studies in genetically diverse populations. *Eur J Hum Genet.* 2012;20(4):469-75.
162. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet.* 2019;20(4):207-20.
163. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nat Med.* 2012;489(7414):57-74.
164. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. *Nat Med.* 2015;518(7539):317-30.
165. ENCODE Project Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.* 2013;45(6):580-5.
166. Klarin D, Emdin CA, Natarajan P, Conrad MF, Kathiresan S. Genetic Analysis of Venous Thromboembolism in UK Biobank Identifies the ZFPM2 Locus and Implicates Obesity as a Causal Risk Factor. *Circ Cardiovasc Genet.* 2017;10(2):e001643.
167. Hinds DA, Buil A, Ziemek D, Martinez-Perez A, Malik R, Folkersen L, et al. Genome-wide association analysis of self-reported events in 6135 individuals and 252 827 controls identifies 8 loci associated with thrombosis. *Hum Mol Genet.* 2016;25(9):1867-74.
168. Bruzelius M, Strawbridge RJ, Tregouet DA, Wiggins KL, Gertow K, Sabater-Lleal M, et al. Influence of coronary artery disease-associated genetic variants on risk of venous thromboembolism. *Thromb Res.* 2014;134(2):426-32.
169. Gregson J, Kaptoge S, Bolton T, Pennells L, Willeit P, Burgess S, et al. Cardiovascular Risk Factors Associated With Venous Thromboembolism. *JAMA Cardiol.* 2019;4(2):163-73.
170. Kim J, Kraft P, Hagan KA, Harrington LB, Lindstroem S, Kabrhel C. Interaction of a genetic risk score with physical activity, physical inactivity, and body mass index in relation to venous thromboembolism risk. *Genet Epidemiol.* 2018;42(4):354-65.

171. Wassel CL, Rasmussen-Torvik LJ, Callas PW, Denenberg JO, Durda JP, Reiner AP, et al. A genetic risk score comprising known venous thromboembolism loci is associated with chronic venous disease in a multi-ethnic cohort. *Thromb Res*. 2015;136(5):966-73.
172. Klok FA, Dzikowska-Diduch O, Kostrubiec M, Vliegen HW, Pruszczyk P, Hasenfuss G, et al. Derivation of a clinical prediction score for chronic thromboembolic pulmonary hypertension after acute pulmonary embolism. *J Thromb Haemost*. 2016;14(1):121-8.
173. Lindstrom S, Germain M, Crous-Bou M, Smith EN, Morange PE, van Hylckama Vlieg A, et al. Assessing the causal relationship between obesity and venous thromboembolism through a Mendelian Randomization study. *Hum Genet*. 2017;136(7):897-902.
174. Satoh T, Satoh K, Yaoita N, Kikuchi N, Omura J, Kurosawa R, et al. Activated TAFI Promotes the Development of Chronic Thromboembolic Pulmonary Hypertension: A Possible Novel Therapeutic Target. *Circ Res*. 2017;120(8):1246-62.
175. Sadler JE. Biochemistry and genetics of von Willebrand factor. *Annu Rev Biochem*. 1998;67:395-424.
176. Dong JF, Moake JL, Nolasco L, Bernardo A, Arceneaux W, Shrimpton CN, et al. ADAMTS-13 rapidly cleaves newly secreted ultralarge von Willebrand factor multimers on the endothelial surface under flowing conditions. *Blood*. 2002;100(12):4033-9.
177. Zheng XL. Structure-function and regulation of ADAMTS-13 protease. *J Thromb Haemost*. 2013;11 Suppl 1:11-23.
178. Baldauf C, Schneppenheim R, Stacklies W, Obser T, Pieconka A, Schneppenheim S, et al. Shear-induced unfolding activates von Willebrand factor A2 domain for proteolysis. *J Thromb Haemost*. 2009;7(12):2096-105.
179. Joly BS, Coppo P, Veyradier A. Thrombotic thrombocytopenic purpura. *Blood*. 2017;129(21):2836-46.
180. Sadler JE. What's new in the diagnosis and pathophysiology of thrombotic thrombocytopenic purpura. *Hematology Am Soc Hematol Educ Program*. 2015;2015:631-6.
181. Rieger M, Mannucci PM, Kremer Hovinga JA, Herzog A, Gerstenbauer G, Konetschny C, et al. ADAMTS13 autoantibodies in patients with thrombotic microangiopathies and other immunomediated diseases. *Blood*. 2005;106(4):1262-7.
182. Mariotte E, Azoulay E, Galicier L, Rondeau E, Zouiti F, Boisseau P, et al. Epidemiology and pathophysiology of adulthood-onset thrombotic microangiopathy with severe ADAMTS13 deficiency (thrombotic thrombocytopenic purpura): a cross-sectional analysis of the French national registry for thrombotic microangiopathy. *Lancet Haematol*. 2016;3(5):e237-45.
183. Terrell DR, Motto DG, Kremer Hovinga JA, Lämmle B, George JN, Vesely SK. Blood group O and black race are independent risk factors for thrombotic thrombocytopenic purpura associated with severe ADAMTS13 deficiency. *Transfusion*. 2011;51(10):2237-43.
184. Rock GA, Shumak KH, Buskard NA, Blanchette VS, Kelton JG, Nair RC, et al. Comparison of plasma exchange with plasma infusion in the treatment of thrombotic thrombocytopenic purpura. Canadian Apheresis Study Group. *New Engl J Med*. 1991;325(6):393-7.
185. Balduini CL, Gugliotta L, Luppi M, Laurenti L, Klersy C, Pieresca C, et al. High versus standard dose methylprednisolone in the acute phase of idiopathic thrombotic thrombocytopenic purpura: a randomized study. *Ann Hematol*. 2010;89(6):591-6.

186. Scully M, McDonald V, Cavenagh J, Hunt BJ, Longair I, Cohen H, et al. A phase 2 study of the safety and efficacy of rituximab with plasma exchange in acute acquired thrombotic thrombocytopenic purpura. *Blood*. 2011;118(7):1746-53.
187. Plaimauer B, Kremer Hovinga JA, Juno C, Wolfsegger MJ, Skalicky S, Schmidt M, et al. Recombinant ADAMTS13 normalizes von Willebrand factor-cleaving activity in plasma of acquired TTP patients by overriding inhibitory antibodies. *J Thromb Haemost*. 2011;9(5):936-44.
188. Levy GG, Nichols WC, Lian EC, Foroud T, McClintick JN, McGee BM, et al. Mutations in a member of the ADAMTS gene family cause thrombotic thrombocytopenic purpura. *Nat Med*. 2001;413(6855):488-94.
189. Pérez-Rodríguez A, Lourés E, Rodríguez-Trillo Á, Costa-Pinto J, García-Rivero A, Batlle-López A, et al. Inherited ADAMTS13 deficiency (Upshaw-Schulman syndrome): a short review. *Thrombosis research*. 2014;134(6):1171-5.
190. Veyradier A, Lavergne JM, Ribba AS, Obert B, Loirat C, Meyer D, et al. Ten candidate ADAMTS13 mutations in six French families with congenital thrombotic thrombocytopenic purpura (Upshaw-Schulman syndrome). *J Thromb Haemost*. 2004;2(3):424-9.
191. Sonneveld MA, de Maat MP, Leebeek FW. Von Willebrand factor and ADAMTS13 in arterial thrombosis: a systematic review and meta-analysis. *Blood Rev*. 2014;28(4):167-78.
192. Tsai AW, Cushman M, Rosamond WD, Heckbert SR, Tracy RP, Aleksic N, et al. Coagulation factors, inflammation markers, and venous thromboembolism: the longitudinal investigation of thromboembolism etiology (LITE). *Am J Med*. 2002;113(8):636-42.
193. Maino A, Siegerink B, Lotta LA, Crawley JT, le Cessie S, Leebeek FW, et al. Plasma ADAMTS-13 levels and the risk of myocardial infarction: an individual patient data meta-analysis. *J Thromb Haemost*. 2015;13(8):1396-404.
194. Mazetto BM, Orsi FL, Barnabe A, De Paula EV, Flores-Nascimento MC, Annichino-Bizzacchi JM. Increased ADAMTS13 activity in patients with venous thromboembolism. *Thromb Res*. 2012;130(6):889-93.
195. Llobet D, Tirado I, Vilalta N, Vallve C, Oliver A, Vazquez-Santiago M, et al. Low ADAMTS13 levels are associated with venous thrombosis risk in women. *Thromb Res*. 2017;157:38-40.
196. Gouvea CP, Matsuda SS, Vaez R, Pinheiro PNB, Noguti MAE, Lourenço DM, et al. The Role Of High Von Willebrand Factor and Low ADAMTS13 Levels In The Risk Of Venous Thromboembolism. *Blood*. 2013;122(21):1128.
197. Domingueti CP, Dusse LM, Fóscolo RB, Reis JS, Annichino-Bizzacchi JM, Orsi FL, et al. Von Willebrand Factor, ADAMTS13 and D-Dimer Are Correlated with Different Levels of Nephropathy in Type 1 Diabetes Mellitus. *PLoS One*. 2015;10(7):e0132784.
198. Konstantinides SV, Torbicki A, Agnelli G, Danchin N, Fitzmaurice D, Galie N, et al. 2014 ESC guidelines on the diagnosis and management of acute pulmonary embolism. *Eur Heart J*. 2014;35(43):3033-69, 69a-69k.
199. Skoro-Sajer N, Gerges C, Gerges M, Panzenböck A, Jakowitsch J, Kurz A, et al. Usefulness of thrombosis and inflammation biomarkers in chronic thromboembolic pulmonary hypertension-sampling plasma and surgical specimens. *J Heart Lung Transplant*. 2018;37(9):1067-74.
200. Orstavik KH, Magnus P, Reisner H, Berg K, Graham JB, Nance W. Factor VIII and factor IX in a twin population. Evidence for a major effect of ABO locus on factor VIII level. *Am J Hum Genet*. 1985;37(1):89-101.

201. Ma Q, Jacobi PM, Emmer BT, Kretz CA, Ozel AB, McGee B, et al. Genetic variants in ADAMTS13 as well as smoking are major determinants of plasma ADAMTS13 levels. *Blood Adv.* 2017;1(15):1037-46.
202. Chion CK, Doggen CJ, Crawley JT, Lane DA, Rosendaal FR. ADAMTS13 and von Willebrand factor and the risk of myocardial infarction in men. *Blood.* 2007;109(5):1998-2000.
203. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet.* 2006;38(2):209-13.
204. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nat Med.* 2007;447(7145):661-78.
205. Bland JM, Altman DG. Statistics notes. The odds ratio. *BMJ.* 2000;320(7247):1468.
206. GenomeStudio Data Analysis Software. Genotyping module v2.0. <https://emea.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html>. Accessed Feb 2017.
207. Guo Y, He J, Zhao S, Wu H, Zhong X, Sheng Q, et al. Illumina human exome genotyping array clustering and quality control. *Nat Protoc.* 2014;9(11):2643-62.
208. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.
209. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. *Nat Med.* 2015;526(7571):68-74.
210. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, et al. Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet.* 2008;83(1):132-5.
211. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics.* 2012;28(24):3326-8.
212. Andersson HM, Siegerink B, Luken BM, Crawley JT, Algra A, Lane DA, et al. High VWF, low ADAMTS13, and oral contraceptives increase the risk of ischemic stroke and myocardial infarction in young women. *Blood.* 2012;119(6):1555-60.
213. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet.* 2011;Chapter 1:Unit1.19.
214. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet.* 2016;48(11):1443-8.
215. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics.* 2014;30(9):1266-72.
216. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet.* 2016;48(10):1279-83.
217. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics.* 2015;31(21):3555-7.
218. Svensson L, Rydberg L, de Mattos LC, Henry SM. Blood group A(1) and A(2) revisited: an immunochemical analysis. *Vox Sang.* 2009;96(1):56-61.

219. Pare G, Chasman DI, Kellogg M, Zee RY, Rifai N, Badola S, et al. Novel association of ABO histo-blood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women. *PLoS Genet.* 2008;4(7):e1000118.
220. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet.* 2007;39(7):906-13.
221. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun.* 2017;8(1):1826.
222. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164.
223. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014;46(3):310-5.
224. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 2012;22(9):1790-7.
225. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods.* 2012;9(3):215-6.
226. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol.* 2015;11(4):e1004219.
227. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics.* 2011;27(12):1739-40.
228. Kokame K, Nobe Y, Kokubo Y, Okayama A, Miyata T. FRET-S-VWF73, a first fluorogenic substrate for ADAMTS13 assay. *Br J Haematol.* 2005;129(1):93-100.
229. Matos LL, Trufelli DC, de Matos MG, da Silva Pinhal MA. Immunohistochemistry as an important tool in biomarkers detection and clinical practice. *Biomark Insights.* 2010;5:9-20.
230. Kim SW, Roh J, Park CS. Immunohistochemistry for Pathologists: Protocols, Pitfalls, and Tips. *J Pathol Transl Med.* 2016;50(6):411-8.
231. de Vries PS, Boender J, Sonneveld MA, Rivadeneira F, Ikram MA, Rottensteiner H, et al. Genetic variants in the ADAMTS13 and SUPT3H genes are associated with ADAMTS13 activity. *Blood.* 2015;125(25):3949-55.
232. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat Commun.* 2017;8:14357.
233. Groemping U. Relative Importance for Linear Regression in R: The Package relaimpo. *Stat Softw.* 2006;17(1):27.
234. Pena EA, Slate EH. gvlma: Global Validation of Linear Models Assumptions. R package version 1.0.0.2. 2014.
235. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2017.
236. Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, Soranzo N, et al. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet.* 2009;5(3):e1000433.
237. McCabe C, White PA, Hoole SP, Axell RG, Priest AN, Gopalan D, et al. Right ventricular dysfunction in chronic thromboembolic obstruction of the pulmonary artery: a pressure-volume study using the conductance catheter. *J Appl Physiol (1985).* 2014;116(4):355-63.

238. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9.
239. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*. 2010;26(18):2336-7.
240. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nat Med*. 2016;536(7616):285-91.
241. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. *Nucleic Acids Res*. 2017;45:D635-42.
242. RStudio Team. RStudio: Integrated Development Environment for R. RStudio, Inc., Boston, MA. 2016.
243. Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4 ed: Springer, New York. 2002.
244. Hothorn T HK, van de Wiel MA, Zeileis A. Implementing a Class of Permutation Tests: The coin Package. *J Stat Softw*. 2008;28(8):23.
245. Pohlert T. The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR). R package. 2014.
246. Long JA. jtools: Analysis and Presentation of Social Scientific Data. R package version 0.9.3. 2017.
247. Kennedy N. forestmodel: Forest Plots from Regression Models. R package version 0.4.3. 2017.
248. Gordon M, Lumley T. forestplot: Advanced Forest Plot Using 'grid' Graphics. R package version 1.9. 2017.
249. Therneau T. A Package for Survival Analysis in S. R package version 2.38. 2015.
250. Kassambara A, Kosinski M. survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.4. 2019.
251. Wickham H. tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. 2017.
252. Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, Kwiatkowski DP, et al. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*. 2007;23(20):2741-6.
253. Schleef M, Strobel E, Dick A, Frank J, Schramm W, Spannagl M. Relationship between ABO and Secretor genotype with plasma levels of factor VIII and von Willebrand factor in thrombosis patients and control individuals. *Br J Haematol*. 2005;128(1):100-7.
254. Rees DC, Cox M, Clegg JB. World distribution of factor V Leiden. *Lancet*. 1995;346(8983):1133-4.
255. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep*. 2016;17(8):2042-59.
256. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*. 2015;348(6235):648-60.
257. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. *Hum Mol Genet*. 2015;24(R1):R111-9.
258. Dichgans M, Malik R, König IR, Rosand J, Clarke R, Gretarsdottir S, et al. Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke*. 2014;45(1):24-36.

259. Heit JA, Armasu SM, Asmann YW, Cunningham JM, Matsumoto ME, Petterson TM, et al. A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. *J Thromb Haemost*. 2012;10(8):1521-31.
260. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet*. 2015;47(10):1121-30.
261. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet*. 2018;50(4):524-37.
262. Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet*. 2016;48(7):709-17.
263. Scott RA, Scott LJ, Magi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes*. 2017;66(11):2888-902.
264. Desch KC, Ozel AB, Siemieniak D, Kalish Y, Shavit JA, Thornburg CD, et al. Linkage analysis identifies a locus for plasma von Willebrand factor undetected by genome-wide association. *Proc Natl Acad Sci U S A*. 2013;110(2):588-93.
265. Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat Genet*. 2010;42(3):210-5.
266. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*. 2016;167(5):1415-29.e19.
267. Qi L, Cornelis MC, Kraft P, Jensen M, van Dam RM, Sun Q, et al. Genetic variants in ABO blood group region, plasma soluble E-selectin levels and risk of type 2 diabetes. *Hum Mol Genet*. 2010;19(9):1856-62.
268. Barbalic M, Dupuis J, Dehghan A, Bis JC, Hoogeveen RC, Schnabel RB, et al. Large-scale genomic studies reveal central role of ABO in sP-selectin and sICAM-1 levels. *Hum Mol Genet*. 2010;19(9):1863-72.
269. Spracklen CN, Chen P, Kim YJ, Wang X, Cai H, Li S, et al. Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels. *Hum Mol Genet*. 2017;26(9):1770-84.
270. Lieb W, Chen MH, Larson MG, Safa R, Teumer A, Baumeister SE, et al. Genome-wide association study for endothelial growth factors. *Circ Cardiovasc Genet*. 2015;8(2):389-97.
271. Pare G, Ridker PM, Rose L, Barbalic M, Dupuis J, Dehghan A, et al. Genome-wide association analysis of soluble ICAM-1 concentration reveals novel associations at the NFKB1K, PNPLA3, RELA, and SH2B3 loci. *PLoS Genet*. 2011;7(4):e1001374.
272. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. *Nat Med*. 2018;558(7708):73-9.
273. Tang W, Schwienbacher C, Lopez LM, Ben-Shlomo Y, Oudot-Mellakh T, Johnson AD, et al. Genetic associations for activated partial thromboplastin time and prothrombin time, their gene expression profiles, and risk of coronary artery disease. *Am J Hum Genet*. 2012;91(1):152-62.
274. Puy C, Rigg RA, McCarty OJ. The hemostatic role of factor XI. *Thromb Res*. 2016;141 Suppl 2:S8-s11.

275. Zhu Z, Yao J, Johns T, Fu K, De Bie I, Macmillan C, et al. SURF1, encoding a factor involved in the biogenesis of cytochrome c oxidase, is mutated in Leigh syndrome. *Nat Genet.* 1998;20(4):337-43.
276. Delcroix M, Vonk Noordegraaf A, Fadel E, Lang I, Simonneau G, Naeije R. Vascular and right ventricular remodelling in chronic thromboembolic pulmonary hypertension. *Eur Respir J.* 2013;41(1):224-32.
277. Standeven KF, Grant PJ, Carter AM, Scheiner T, Weisel JW, Ariëns RA. Functional analysis of the fibrinogen Aalpha Thr312Ala polymorphism: effects on fibrin structure and function. *Circulation.* 2003;107(18):2326-30.
278. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gómez-Marín C, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nat Med.* 2014;507(7492):371-5.
279. Groot HE, Villegas Sierra LE, Said MA, Lipsic E, Karper JC, van der Harst P. Genetically Determined ABO Blood Group and its Associations With Health and Disease. *Arterioscler Thromb Vasc Biol.* 2020;40(3):830-8.
280. Dean L. Medical Genetics Summaries: ABO blood group. [Online]. [Accessed 1 March 2020] Available from: <https://www.ncbi.nlm.nih.gov/books/NBK100894/>.
281. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. *Nat Methods.* 2011;8(10):833-5.
282. Zaykin DV, Kozbur DO. P-value based analysis for shared controls design in genome-wide association studies. *Genet Epidemiol.* 2010;34(7):725-38.
283. Crawley JT, Lam JK, Rance JB, Mollica LR, O'Donnell JS, Lane DA. Proteolytic inactivation of ADAMTS13 by thrombin and plasmin. *Blood.* 2005;105(3):1085-93.
284. Feys HB, Vandeputte N, Palla R, Peyvandi F, Peerlinck K, Deckmyn H, et al. Inactivation of ADAMTS13 by plasmin as a potential cause of thrombotic thrombocytopenic purpura. *J Thromb Haemost.* 2010;8(9):2053-62.
285. Moake JL, Rudy CK, Troll JH, Weinstein MJ, Colannino NM, Azocar J, et al. Unusually large plasma factor VIII: von Willebrand factor multimers in chronic relapsing thrombotic thrombocytopenic purpura. *N Engl J Med.* 1982;307(23):1432-5.
286. Schwameis M, Schorzenhofer C, Assinger A, Steiner MM, Jilma B. VWF excess and ADAMTS13 deficiency: a unifying pathomechanism linking inflammation to thrombosis in DIC, malaria, and TTP. *Thromb Haemost.* 2015;113(4):708-18.
287. Wolff B, Lodziewski S, Bollmann T, Opitz CF, Ewert R. Impaired peripheral endothelial function in severe idiopathic pulmonary hypertension correlates with the pulmonary vascular response to inhaled iloprost. *Am Heart J.* 2007;153(6):1088.e1-7.
288. Schäfer M, Kheyfets VO, Schroeder JD, Dunning J, Shandas R, Buckner JK, et al. Main pulmonary arterial wall shear stress correlates with invasive hemodynamics and stiffness in pulmonary hypertension. *Pulm Circ.* 2016;6(1):37-45.
289. South K, Freitas MO, Lane DA. A model for the conformational activation of the structurally quiescent metalloprotease ADAMTS13 by von Willebrand factor. *J Biol Chem.* 2017;292(14):5760-9.
290. Zhao BQ, Chauhan AK, Canault M, Patten IS, Yang JJ, Dockal M, et al. von Willebrand factor-cleaving protease ADAMTS13 reduces ischemic brain injury in experimental stroke. *Blood.* 2009;114(15):3329-34.

291. De Meyer SF, Savchenko AS, Haas MS, Schatzberg D, Carroll MC, Schiviz A, et al. Protective anti-inflammatory effect of ADAMTS13 on myocardial ischemia/reperfusion injury in mice. *Blood*. 2012;120(26):5217-23.
292. Witsch T, Martinod K, Sorvillo N, Portier I, De Meyer SF, Wagner DD. Recombinant Human ADAMTS13 Treatment Improves Myocardial Remodeling and Functionality After Pressure Overload Injury in Mice. *J Am Heart Assoc*. 2018;7(3):e007004.
293. Gandhi C, Motto DG, Jensen M, Lentz SR, Chauhan AK. ADAMTS13 deficiency exacerbates VWF-dependent acute myocardial ischemia/reperfusion injury in mice. *Blood*. 2012;120(26):5224-30.
294. Bonderman D, Jakowitsch J, Adlbrecht C, Schemper M, Kyrle PA, Schonauer V, et al. Medical conditions increasing the risk of chronic thromboembolic pulmonary hypertension. *Thromb Haemost*. 2005;93(3):512-6.
295. Chauhan AK, Kisucka J, Brill A, Walsh MT, Scheiflinger F, Wagner DD. ADAMTS13: a new link between thrombosis and inflammation. *J Exp Med*. 2008;205(9):2065-74.
296. Lee M, Rodansky ES, Smith JK, Rodgers GM. ADAMTS13 promotes angiogenesis and modulates VEGF-induced angiogenesis. *Microvasc Res*. 2012;84(2):109-15.
297. Xu H, Cao Y, Yang X, Cai P, Kang L, Zhu X, et al. ADAMTS13 controls vascular remodeling by modifying VWF reactivity during stroke recovery. *Blood*. 2017;130(1):11-22.
298. Lotta LA, Tuana G, Yu J, Martinelli I, Wang M, Yu F, et al. Next-generation sequencing study finds an excess of rare, coding single-nucleotide variants of ADAMTS13 in patients with deep vein thrombosis. *J Thromb Haemost*. 2013;11(7):1228-39.
299. Freeman WM, Walker SJ, Vrana KE. Quantitative RT-PCR: pitfalls and potential. *Biotechniques*. 1999;26(1):112-22, 24-5.
300. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57-63.
301. Folsom AR, Tang W, Weng LC, Roetker NS, Cushman M, Basu S, et al. Replication of a genetic risk score for venous thromboembolism in whites but not in African Americans. *J Thromb Haemost*. 2016;14(1):83-8.
302. de Haan HG, Bezemer ID, Doggen CJ, Le Cessie S, Reitsma PH, Arellano AR, et al. Multiple SNP testing improves risk prediction of first venous thrombosis. *Blood*. 2012;120(3):656-63.
303. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. *Genomics Inform*. 2012;10(2):117-22.
304. Kujovich JL. Factor V Leiden thrombophilia. *Genet Med*. 2011;13(1):1-16.
305. Kyrle PA, Rosendaal FR, Eichinger S. Risk assessment for recurrent venous thrombosis. *Lancet*. 2010;376(9757):2032-9.
306. Pomero F, Ageno W, Serraino C, Borretta V, Gianni M, Fenoglio L, et al. The role of inherited thrombophilia in patients with isolated pulmonary embolism: a systematic review and a meta-analysis of the literature. *Thromb Res*. 2014;134(1):84-9.
307. de Moerloose P, Reber G, Perrier A, Perneger T, Bounameaux H. Prevalence of factor V Leiden and prothrombin G20210A mutations in unselected patients with venous thromboembolism. *Br J Haematol*. 2000;110(1):125-9.

308. Corral J, Roldán V, Vicente V. Deep venous thrombosis or pulmonary embolism and factor V Leiden: enigma or paradox. *Haematologica*. 2010;95(6):863-6.
309. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics*. 2014;13(2):397-406.
310. Therrien M, Chang HC, Solomon NM, Karim FD, Wassarman DA, Rubin GM. KSR, a novel protein kinase required for RAS signal transduction. *Cell*. 1995;83(6):879-88.
311. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet*. 2015;47(9):979-86.
312. Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J, Han B, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat Genet*. 2016;48(5):510-8.
313. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42:D980-5.
314. Kratochvílová K, Horak P, Ešner M, Souček K, Pils D, Anees M, et al. Tumor suppressor candidate 3 (TUSC3) prevents the epithelial-to-mesenchymal transition and inhibits tumor growth by modulating the endoplasmic reticulum stress response in ovarian cancer cells. *Int J Cancer*. 2015;137(6):1330-40.
315. Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, et al. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet*. 2019;104(1):65-75.
316. Zhang Q, Marioni RE, Robinson MR, Higham J, Sproul D, Wray NR, et al. Genotype effects contribute to variation in longitudinal methylome patterns in older people. *Genome Med*. 2018;10(1):75.
317. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet*. 2019;20(8):467-84.
318. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. *Nat Rev Genet*. 2009;10(10):681-90.
319. Ten Cate V, Koeck T, Panova-Noeva M, Rapp S, Prochaska JH, Lenz M, et al. A prospective cohort study to identify and evaluate endotypes of venous thromboembolism: Rationale and design of the Genotyping and Molecular Phenotyping in Venous ThromboEmbolic project (GMP-VTE). *Throm Res*. 2019;181:84-91.
320. Yu S, Kumamaru KK, George E, Bedayat A, Rybicki FJ, Dunne RM, et al. Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing. *J Biomed Inform*. 2014;52:386-93.
321. Cho JH, Feldman M. Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies. *Nat Med*. 2015;21(7):730-8.
322. Rahaghi FN, Argemí G, Nardelli P, Domínguez-Fandos D, Arguis P, Peinado VI, et al. Pulmonary vascular density: comparison of findings on computed tomography imaging with histology. *Eur Respir J*. 2019;54(2):1900370.

323. Rahaghi FN, Ross JC, Agarwal M, González G, Come CE, Diaz AA, et al. Pulmonary vascular morphology as an imaging biomarker in chronic thromboembolic pulmonary hypertension. *Pulm Circ.* 2016;6(1):70-81.
324. Dahl A, Cai N, Ko A, Laakso M, Pajukanta P, Flint J, et al. Reverse GWAS: Using genetics to identify and model phenotypic subtypes. *PLoS genetics.* 2019;15(4):e1008009.
325. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med.* 2015;7(311):311ra174.
326. Zakai NA, McClure LA. Racial differences in venous thromboembolism. *J Thromb Haemost.* 2011;9(10):1877-82.
327. Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front Genet.* 2020;11:424.
328. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. Opportunities and challenges for transcriptome-wide association studies. *Nat Genet.* 2019;51(4):592-9.
329. Ulrich A, Wharton J, Thayer TE, Swietlik EM, Assad TR, Desai AA, et al. Mendelian randomisation analysis of red cell distribution width in pulmonary arterial hypertension. *Eur Respir J.* 2020;55(2): 1901486.
330. Malik N, Rao MS. A review of the methods for human iPSC derivation. *Methods Mol Biol.* 2013;997:23-33.
331. Coll M, Perea L, Boon R, Leite SB, Vallverdú J, Mannaerts I, et al. Generation of Hepatic Stellate Cells from Human Pluripotent Stem Cells Enables In Vitro Modeling of Liver Fibrosis. *Cell Stem Cell.* 2018;23(1):101-13.e7.
332. Adams WJ, Zhang Y, Cloutier J, Kuchimanchi P, Newton G, Sehrawat S, et al. Functional vascular endothelium derived from human induced pluripotent stem cells. *Stem Cell Reports.* 2013;1(2):105-13.
333. Mercier O, Tivane A, Dorfmueller P, de Perrot M, Raoux F, Decante B, et al. Piglet model of chronic pulmonary hypertension. *Pulm Circ.* 2013;3(4):908-15.
334. Nichols TC, Bellinger DA, Merricks EP, Raymer RA, Kloos MT, Defriess N, et al. Porcine and canine von Willebrand factor and von Willebrand disease: hemostasis, thrombosis, and atherosclerosis studies. *Thrombosis.* 2010;2010:461238.
335. Fujioka M, Hayakawa K, Mishima K, Kunizawa A, Irie K, Higuchi S, et al. ADAMTS13 gene deletion aggravates ischemic brain damage: a possible neuroprotective role of ADAMTS13 by ameliorating postischemic hypoperfusion. *Blood.* 2010;115(8):1650-3.
336. Sonneveld MA, de Maat MP, Portegies ML, Kavousi M, Hofman A, Turecek PL, et al. Low ADAMTS13 activity is associated with an increased risk of ischemic stroke. *Blood.* 2015;126(25):2739-46.
337. Banno F, Kokame K, Okuda T, Honda S, Miyata S, Kato H, et al. Complete deficiency in ADAMTS13 is prothrombotic, but it alone is not sufficient to cause thrombotic thrombocytopenic purpura. *Blood.* 2006;107(8):3161-6.
338. Denorme F, Langhauser F, Desender L, Vandenbulcke A, Rottensteiner H, Plaimauer B, et al. ADAMTS13-mediated thrombolysis of t-PA-resistant occlusions in ischemic stroke in mice. *Blood.* 2016;127(19):2337-45.
339. Bustamante A, Ning M, García-Berrocso T, Penalba A, Boada C, Simats A, et al. Usefulness of ADAMTS13 to predict response to recanalization therapies in acute ischemic stroke. *Neurology.* 2018;90(12):e995-e1004.

Appendix

Below are the listed copyright permissions for figures reproduced in this thesis.

Figure 1.1

Reproduced with permission of the © ERS 2020: European Respiratory Review 2015; 24: 263-271; DOI: 10.1183/16000617.00000815 (Figure 1)

Permission from Elin Reeves, Director of Publications, European Respiratory Society, 15th July 2020.

Figure 1.2

Reproduced with permission of the © Elsevier: Seminars in Thoracic and Cardiovascular Surgery 2006; 18(3): 243-249 (Figure 4)

Permission and license via the Copyright Clearance Center, license number 4857220717385, 27th June 2020. For reuse in thesis: electronic and print.

Figure 1.3

Reproduced under creative commons license: [CC BY-SA 3.0](https://creativecommons.org/licenses/by-sa/3.0/)

Figure 1.4

Reproduced under creative commons license: [CC BY-SA 3.0](https://creativecommons.org/licenses/by-sa/3.0/)

Figure 1.5

Reproduced with permission of the © CMAJ group (www.cmaj.ca): Canadian Medical Association Journal 2002; 167(1): 48-54 (Figure 1)

Permission from Holly Bodger, Publisher, The CMAJ Group, 15th July 2020.

Figure 1.6

Reproduced with permission of the © BMJ Publishing Group Ltd: The British Medical Journal 2014; 349: g4387 (Figure on page 9)

Permission and license via the Copyright Clearance Center, license number 4857230439576, 27th June 2020. For reuse in thesis: electronic and print.

Figure 1.7

Reproduced with permission of the © Springer Nature: Nature Reviews Cardiology 2020; 17(2): 85-95 (Figure 1)

Permission and license via the Copyright Clearance Center, license number 4870230134454, 15th July 2020. For reuse in thesis: electronic and print.

Figure 1.8

Reproduced with permission of the Copyright: © 2012 Bush, Moore: PLoS computational biology 2012; 8(12):e1002822 (Figure 1)

This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Figure 1.9

Reproduced with permission of the © Elsevier: The American Journal of Human Genetics 2017; 101(1):5-22 (Figure 2)

Permission and license via the Copyright Clearance Center, license number 4870270384354, 15th July 2020. For reuse in thesis: electronic and print.

Figure 1.10

Reproduced with permission of the © Illumina:

<https://dnatech.genomecenter.ucdavis.edu/infinium-assay/> (Figure 1)

Permission from, Rufus Knight, Direct Marketing Specialist (EMEA), Illumina, 16th July 2020.

Figure 1.11

Reproduced with permission of the © Massachusetts Medical Society: The New England Journal of Medicine 2010; 363(2): 166-76 (Figure 1)

Content (full-text or portions thereof) may be used in print and electronic versions of a dissertation or thesis without formal permission from the Massachusetts Medical Society (MMS), Publisher of the New England Journal of Medicine.

Figure 1.12

Reproduced with permission of the Copyright: © 2012 Bush, Moore: PLoS computational biology 2012; 8(12):e1002822 (Figure 2)

This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Figure 1.13

Reproduced with permission of the © Springer Nature: Nature Reviews Genetics 2010; 11(7): 499-511 (Figure in Box 1)

Permission and license via the Copyright Clearance Center, license number 4870271354233, 15th July 2020. For reuse in thesis: electronic and print.

Figure 1.14

Reproduced with permission of the © Elsevier: The American Journal of Human Genetics 2015; 96(4): 532-42 (Figure S1)

Permission and license via the Copyright Clearance Center, license number 4870280118598, 15th July 2020. For reuse in thesis: electronic and print.

Figure 1.15

This research was originally published in Blood. Joly BS, Coppo P, Veyradier A. Thrombotic thrombocytopenic purpura. Blood. 2017;129(21):2836-46 (Figure 2)

© the American Society of Hematology